ATTENTION INVARIANTS: WEIGHT-SPACE INTERPRETABILITY WITHOUT A FORWARD PASS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention heads in transformers can be characterized through basis-invariant quantities from weights alone, without any data or forward pass. We derive two invariants of the QK sector: ρ, which measures a head's positional capacity, and α, the definiteness of the symmetric part of its QK form. We prove that attention heads obey a necessity condition on positional computation. A head with $\rho = 0$ cannot depend on relative position, so positional function requires $\rho > 0$, a structural constraint we find is obeyed without exception across every model. On five models from four architecture families (Qwen2.5-1.5B/7B, Mistral-7B, Llama-3.1-8B, Gemma-2-2B), ρ and α separate three canonical head types (induction, duplicate-token, previous-token) in distinct regions of a two-dimensional plane. We show that a logistic regression trained on (ρ, α) separates every head type pair at ROC-AUC ≥ 0.97, and unsupervised clustering recovers the types at 0.86–1.00 purity. Transferring to a different architecture, we obtain a median of 0.94 balanced accuracy (16 of 20 pairs ≥ 0.87). As an application, we apply YaRN only to the top 50% heads by ρ, arriving at better results than a standard uniform application.

1 INTRODUCTION

The widespread adoption of Large Language Models (LLMs) calls for a deeper understanding of the inner workings of Transformers (Vaswani et al., 2017). Works on mechanistic interpretability are generally focused on the behaviour of attention head activations (alongside the papers cited in Section 2, see for example Ameisen et al. (2025), Ferrando et al. (2023), Sakarvadia et al. (2023)). But what can we learn from the model weights?

It was previously shown in the literature that weights in an LLM store factual knowledge (Meng et al., 2022) and attention head functionality can be inferred from them (Elhelo & Geva, 2025). Here, we further uncover information that model weights can provide us by looking into the symmetries of attention heads. The parameter spaces where the d-dimensional Query-Key (QK) and Value-Output (VO) projections live are invariant under orthogonal transformations, yielding a (continuous) $O(d)$ symmetry group. Rotary Positional Encoding (RoPE; Su et al. (2024)), which is applied in most state-of-the-art LLMs, act on the QK space by partitioning query and key vectors into $d/2$ two-dimensional components and rotating each component. Therefore, the symmetry of the QK space is broken down from $O(d)$ to $SO(2)^{d/2}$. The invariance of the QK component of the attention head under the residual $SO(2)^{d/2}$ symmetry makes it a basis-independent observable of the head's structure. This allows us to derive scalar quantities that characterize attention heads independently of the choice of basis. Two such invariants are presented in this paper, namely ρ, which provides a measure for the positional capacity of a head, and α, which can be interpreted as content-matching definiteness.

The invariants ρ and α identify structural features of a head directly from its weights, requiring neither data nor forward passes. This is complementary to activation-based interpretability, and can strengthen it. The invariants give a cheap, architecture independent coordinate system that can guide and narrow the more expensive analyses that require forward passes. Our main goal in this paper is to establish this weights-only characterization and to show that it recovers, and transfers across models, the structure of known head types.

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

This paper is structured as follows: In Section 2 we mention related work that inspired our paper. Next, in Section 3.1, we describe in more detail the symmetries of attention, as well as the invariants ρ and α. After Section 4 where we present our method to unveil the properties of attention heads, we show some of our main results in Section 5. Specifically, we make use of supervised (logistic regression) and unsupervised (k-means clustering) techniques to show that ρ and α together can be used to classify three canonical head types, induction, duplicate-token and previous-token, in different models. This classification provided by the two invariants is also shown to be transferrable across different architectures with median balanced accuracy of 0.94. In Section 6, an application of the invariant ρ to capacity-gated context extension is presented. The current limitations of our work, which we intend to address in the future, are discussed in Section 7. Finally, some concluding remarks are made in Section 8.

2 Related Work

Symmetry group invariants. Discovering observable parameters in a model through the search for symmetry group invariants is common in geometric deep learning (Maron et al., 2019) and in natural sciences such as physics. This paper takes inspiration from those fields, deriving weight-only observables from the basis symmetry of the QK sector of attention heads.

Characterization of attention heads. A large body of work identifies functional head types from their behavior on data. The most studied example are induction heads, which implement the "[A][B]…[A]→[B]" copying rule (Elhage et al., 2021; Olsson et al., 2022). Other head types explored in the literature are previous-token and duplicate-token heads from the same induction-circuit analysis, name-mover and S-inhibition heads in the indirect-object-identification circuit (Wang et al., 2023), successor heads (Gould et al., 2024), and copy-suppression heads (McDougall et al., 2023). Another research area consists on the automation of the search for those head types, with the method of (Conmy et al., 2023) being an example of how such task may be performed. The aforementioned head characterizations require forward passes on curated inputs and read out activations, whereas ours requires no data and no forward pass.

RoPE and the positional/content distinction. As we described in Section 1, RoPE rotates two-dimensional query and key subvectors. This gives the model the ability to capture relative token distances without the computational overhead of modifying the attention matrix directly. Barbero et al. (2024) show that heads use high RoPE frequencies for robust positional patterns and low frequencies for semantic "channels", while Chen & Yan (2024) relate QK weight angles to whether a head is position- or semantics-sensitive. Rather than observing multiple rotation angles in 2D-planes, we are able to use a single parameter — the invariant ρ — to measure how much a head can couple its content interaction to relative position.

Positional vs. symbolic behavior. Our work has a close resemblance with Urrutia et al. (2026b), who formalize a positional↔symbolic dichotomy of heads and quantify it with a score computed from attention weights on a task (permutation invariance/equivariance of the logits on a binding task). However, their score and ρ differ fundamentally in method and object, even though the two quantities are closely related in terms of concept. In fact, the score defined in Urrutia et al. (2026b) is a property of a (head, input) pair obtained from a forward pass, whereas ρ is a property of the head itself, computed from weights without data. In other words, they measure whether a head behaves positionally on a given input, while we measure a head's structural capacity to do so. Furthermore, the invariant α that we also present has no analogue in their framework: it resolves structure within the content regime, placing heads on a retrieval↔aggregation axis, which a single positional/symbolic score cannot express. Two concurrent works reinforce that the position/content distinction that ρ measures is functionally load-bearing. Urrutia et al. (2026a) study the learning dynamics of positional versus symbolic heads and show that symbolic mechanisms extrapolate to longer sequences more reliably than positional ones. This is a behavioral counterpart to the capacity that ρ reads from weights. Gopalakrishnan et al. (2026) argue that "what" (content) and "where" (position) are entangled in RoPE and propose an improvement to RoPE dubbed Polar Coordinate Position Embedding (PoPE). Our decomposition $\boldsymbol{G}_{\mathrm{QK}}(\Delta) = \boldsymbol{G}_{\mathrm{QK}}(0) + \boldsymbol{P}(\Delta)$ (see Section 3) separates the same two factors on the measurement side, quantifying per head the coupling they seek to remove by construction.

Context extension. Positional interpolation (Chen et al., 2023) and *Yet another RoPE extensioN method* (YaRN; Peng et al. (2024)) extend context by rescaling RoPE uniformly across all heads. The base frequency itself bounds attainable context (Men et al., 2024), at the same time positional encoding shapes length generalization (Kazemnejad et al., 2023). In Section 6, we use ρ to selectively apply such scaling to position-dominated heads.

3 Theory: invariants of the attention head

3.1 Symmetries of attention

Inside a single head, the query and key projections $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_h \times d}$ map the d-dimensional residual stream to a d_h-dimensional head space, acting on residual-stream vectors $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d$ at query position i and key position j. RoPE multiplies queries and keys by a position-dependent rotation matrix $\boldsymbol{R}(m) = \bigoplus_{k=1}^{d_h/2} \mathrm{Rot}(\omega_k m) \in SO(2)^{d_h/2}$, block-diagonal in the $d_h/2$ frequency planes with frequencies ω_k, before the inner product. Writing $\Delta = i - j$ for the relative position and using $\boldsymbol{R}(i)^\top \boldsymbol{R}(j) = \boldsymbol{R}(-\Delta)$, the pre-softmax attention logit is the bilinear form

$$s_{ij} = \boldsymbol{x}_i^\top \boldsymbol{G}_{\mathrm{QK}}(\Delta)\, \boldsymbol{x}_j, \qquad \boldsymbol{G}_{\mathrm{QK}}(\Delta) = \boldsymbol{W}_Q^\top \boldsymbol{R}(\Delta) \boldsymbol{W}_K \in \mathbb{R}^{d \times d}. \tag{1}$$

The head admits two exact symmetries that leave the logits invariant: a rotation $\boldsymbol{S} \in SO(2)^{d_h/2}$

$$(\boldsymbol{S}\boldsymbol{W}_Q)^\top \left(\boldsymbol{S}\boldsymbol{R}(\Delta)\boldsymbol{S}^\top\right) (\boldsymbol{S}\boldsymbol{W}_K) = \boldsymbol{G}_{\mathrm{QK}}(\Delta), \tag{2}$$

which commutes with $\boldsymbol{R}(\Delta)$, and a softmax shift symmetry (c.f. Appendix C). Any observable of this head must be invariant under both symmetries; in this paper we concern ourselves with the former.

3.2 The ρ invariant

For models using RoPE, the attention logit decomposes as

$$s_{ij} = \boldsymbol{x}_i^\top \boldsymbol{G}_{\mathrm{QK}}(\Delta)\, \boldsymbol{x}_j = \boldsymbol{x}_i^\top \boldsymbol{G}_{\mathrm{QK}}(0)\, \boldsymbol{x}_j + \boldsymbol{x}_i^\top \boldsymbol{P}(\Delta)\, \boldsymbol{x}_j, \tag{3}$$

where $\boldsymbol{G}_{\mathrm{QK}}(0) = \boldsymbol{W}_Q^\top \boldsymbol{W}_K$ is the static term and $\boldsymbol{P}(\Delta) = \boldsymbol{W}_Q^\top [\boldsymbol{R}(\Delta) - \boldsymbol{I}] \boldsymbol{W}_K$ is the positional perturbation at relative position $\Delta = i - j$. We define

$$\rho = \frac{1}{|\mathcal{D}|} \sum_{\Delta \in \mathcal{D}} \frac{\|\boldsymbol{P}(\Delta)\|_F^2}{\|\boldsymbol{G}_{\mathrm{QK}}(0)\|_F^2}, \qquad \mathcal{D} = \{2^n\}_{n=0}^{\lfloor \log_2 L \rfloor}. \tag{4}$$

The invariant ρ has four properties that make it a well-posed observable of the head, all directly from equation 3–equation 4. Non-negativity, $\rho \geq 0$, with $\rho = 0$ iff $\boldsymbol{R}(\Delta)\boldsymbol{W}_K = \boldsymbol{W}_K$ for all $\Delta \in \mathcal{D}$. Scale invariant, as ρ is unchanged under $\boldsymbol{W}_Q \to c\boldsymbol{W}_Q$ or $\boldsymbol{W}_K \to c\boldsymbol{W}_K$. Additionally, because $\boldsymbol{R}(\Delta)$ is block-diagonal across the $d_h/2$ frequency planes, the perturbation splits as $\boldsymbol{P}(\Delta) = \sum_{k=1}^{d_h/2} \boldsymbol{W}_Q^\top [\boldsymbol{R}_k(\Delta) - \boldsymbol{I}_k] \boldsymbol{W}_K$, where $\boldsymbol{R}_k$ acts only on plane k. Thus,

$$\rho = \sum_{k=1}^{d_h/2} \rho_k, \qquad \rho_k = \frac{1}{|\mathcal{D}|} \sum_{\Delta \in \mathcal{D}} \frac{\|\boldsymbol{P}_k(\Delta)\|_F^2}{\|\boldsymbol{G}_{\mathrm{QK}}(0)\|_F^2}. \tag{5}$$

This is the weights-only, basis-invariant counterpart of the empirical observation that heads allocate distinct RoPE frequencies to positional versus semantic use (Barbero et al., 2024; Chen & Yan, 2024). Finally, ρ is invariant under the residual symmetry $SO(2)^{d_h/2}$ of equation 2, and the Frobenius norm allows a scalar measure for positional capacity.

We stress that as ρ is a property of the weights, it quantifies the capacity of a head to couple its content interaction to relative position, independent of any input. A head with large ρ is built to be able to read position, regardless of using this property on a particular input. Conversely, a head with small ρ does not have the capacity to use positional information easily. Thus, ρ is a data-free measure of head function rather than a behavioral measurement (cf. activation-based positional/symbolic scores, which are properties of a head and input pair (Urrutia et al., 2026b)). We return to this distinction in Section 7.

Proposition 1 ($\rho = 0$ forbids positional encoding)**.** *Fix a head with static term $\boldsymbol{G}_{\mathrm{QK}}(0) = \boldsymbol{W}_Q^\top \boldsymbol{W}_K$ and let $\mathcal{D}$ be any set of relative positions. Then $\rho = 0$ if and only if $\boldsymbol{P}(\Delta) = \mathbf{0}$ for every $\Delta \in \mathcal{D}$, in which case the attention logit*

$$s_{ij} = \boldsymbol{x}_i^\top \boldsymbol{G}_{\mathrm{QK}}(\Delta)\, \boldsymbol{x}_j = \boldsymbol{x}_i^\top \boldsymbol{G}_{\mathrm{QK}}(0)\, \boldsymbol{x}_j \tag{6}$$

is independent of the relative position $\Delta = i - j$ for all $\Delta \in \mathcal{D}$ and all residual-stream inputs $\boldsymbol{x}_i, \boldsymbol{x}_j$.

Proof. By equation 4, ρ is a sum of terms $\|\boldsymbol{P}(\Delta)\|_F^2 / \|\boldsymbol{G}_{\mathrm{QK}}(0)\|_F^2$, each a squared Frobenius norm over a common positive denominator, hence each ≥ 0. A sum of non-negative terms vanishes iff every term vanishes, so $\rho = 0$ iff $\boldsymbol{P}(\Delta) = \mathbf{0}$ for every $\Delta \in \mathcal{D}$. Substituting $\boldsymbol{P}(\Delta) = \boldsymbol{G}_{\mathrm{QK}}(\Delta) - \boldsymbol{G}_{\mathrm{QK}}(0) = \mathbf{0}$ into equation 3 gives $s_{ij} = \boldsymbol{x}_i^\top \boldsymbol{G}_{\mathrm{QK}}(0)\, \boldsymbol{x}_j$, which contains no dependence on Δ. □

A head with $\rho = 0$ therefore assigns to any content pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ the same score at every relative position. It cannot implement any function whose output depends on where a token sits relative to another. Positional function is not merely unused at $\rho = 0$, it cannot exist.

3.3 THE α INVARIANT

Using the remaining $\boldsymbol{G}_{\mathrm{QK}}(0)$ term in equation 3, we may construct a second invariant. Any square matrix decomposes uniquely into symmetric and antisymmetric parts. We may then define,

$$\boldsymbol{G}_{\mathrm{QK}}(0) = \boldsymbol{G}_+ + \boldsymbol{G}_-, \qquad \boldsymbol{G}_\pm = \tfrac{1}{2}\big(\boldsymbol{G}_{\mathrm{QK}}(0) \pm \boldsymbol{G}_{\mathrm{QK}}(0)^\top\big)\,. \tag{7}$$

The antisymmetric part $\boldsymbol{G}_-$ measures directionality of the content interaction, i.e. the extent to which the score $\boldsymbol{x}_i^\top \boldsymbol{G}_{\mathrm{QK}}(0)\, \boldsymbol{x}_j$ differs from $\boldsymbol{x}_j^\top \boldsymbol{G}_{\mathrm{QK}}(0)\, \boldsymbol{x}_i$. Under the causal mask, where only $\Delta = i - j \geq 0$ is ever evaluated, this component is weakly constrained, and we restrict our attention to the symmetric part $\boldsymbol{G}_+$.

Since $\boldsymbol{G}_+$ is real symmetric, it has real eigenvalues $\lambda_1, \ldots, \lambda_d$ and its signed definiteness is a basis-independent scalar. Thus, we can define

$$\alpha = \frac{\mathrm{Tr}(\boldsymbol{G}_+)}{\|\boldsymbol{G}_+\|_*} = \frac{\sum_k \lambda_k}{\sum_k |\lambda_k|} \;\in\; [-1, 1], \tag{8}$$

where $\|\cdot\|_*$ is the nuclear norm. We note that α is scale-invariant: it is unchanged under $\boldsymbol{G}_+ \to c\boldsymbol{G}_+$. To interpret α we notice that it is controlled by the definiteness of $\boldsymbol{G}_+$. A consequence can be promptly seen in the diagonal of $[\boldsymbol{G}_{\mathrm{QK}}(0)]_{kk} = [\boldsymbol{G}_+]_{kk}$. The definiteness of $\boldsymbol{G}_+$ controls the sign of the self-interaction between key feature k and query feature k. As such:

- $\alpha \to +1$: $\boldsymbol{G}_+$ is (near) positive semi-definite. Every self-interaction $(\boldsymbol{G}_+)_{kk} \geq 0$, and by positivity of the 2×2 principal minors every cross-feature coupling is Cauchy-Schwarz bounded by the self-couplings,$(\boldsymbol{G}_+)_{jk}^2 \leq (\boldsymbol{G}_+)_{jj}(\boldsymbol{G}_+)_{kk}$. The form is thus a similarity kernel dominated by its diagonal. The head scores highest for keys that resemble the query, as we would expect in a duplicate-token head;
- $\alpha \to -1$: $\boldsymbol{G}_+$ is (near) negative semi-definite, so every self-interaction $(\boldsymbol{G}_+)_{kk} \leq 0$. The resemblance to the query is scored negatively;
- $\alpha \approx 0$: $\boldsymbol{G}_+$ is indefinite, with eigenvalues of both signs in comparable proportion. The form is then neither a similarity nor an anti-similarity kernel.

Thus, α measures the definiteness of the content interaction $\boldsymbol{G}_+$, interpolating between a similarity kernel ($\alpha \to +1$) and an anti-similarity kernel ($\alpha \to -1$). By construction it is built from a disjoint part of $\boldsymbol{G}_{\mathrm{QK}}$ than ρ, being a function of the static term $\boldsymbol{G}_{\mathrm{QK}}(0)$ alone. More precisely, ρ and α are the leading invariants of the two sectors that the residual symmetry separates, the position-dependent part and the static term. Appendix B counts the basis-independent degrees of freedom of the QK sector and places both invariants within them.

4 METHOD

Our results in the next section are presented for five models: Qwen-1.5B and Qwen-7B (Qwen Team, 2025), Mistral-7B (Jiang et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024) and Gemma-2-2B (Gemma Team, 2024). For each model, we extract the weight matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K$ per head, explicitly handling Group-Query Attention (GQA) mechanisms and variable head dimensions (`head_dim`). RoPE frequencies are read from the model's own `inv_freq`. With the parameters mentioned above, the invariants ρ and α are computed without any forward pass.

The behavioral labels, which classify each attention head as *induction*, *duplicate* or *previous-token* (Olsson et al., 2022; Elhage et al., 2021), are collected from repeated sequences of the form $[\text{BOS}, r, r]$, where r is a block of 50 random tokens repeated once. On such a sequence, each head type has an attention pattern. A previous-token head attends from position i to $i-1$, a duplicate-token head attends from a token in the second copy back to its first occurrence, and an induction head attends from a token in the second copy to the token that followed its first occurrence. For each head we measure the average attention mass on this defining pattern, assign the dominant type when that mass exceeds a threshold (we use 0.30), and label the head "other" otherwise. The resulting label counts are small, as named heads are rare. We report them per model in Appendix E. Appendix D shows the results are stable across a range of thresholds.

For supervised separability and cross-model transfer, we use ℓ_2 logistic regression with standardized features and balanced class weights. The two features are $\log(1+\rho)$ and α, the log being a monotonic rescaling that leaves the ordering of heads by ρ unchanged. For separability, we fit a classifier to each pair of head types within a model and report ROC-AUC. For transfer, we fit a three-class classifier on the labeled heads of one model, predict on another, and report balanced accuracy (chance $1/3$). We also perform unsupervised recovery with k-means clustering ($k = 3$), evaluated by label purity. We note that a two-feature model has little capacity to overfit, and as such its performance should reflect the structure in ρ and α and not a property of the classifier.

5 RESULTS

5.1 HEAD TYPES SEPARATE IN THE (ρ, α) PLANE

Figure 1 shows every head of each model placed on the (ρ, α) plane. The labeled heads fall into three regions that are the same in all five models. Induction heads sit at low ρ with a small positive α, duplicate-token heads at low ρ with high α, and previous-token heads at high ρ. These placements are what the invariants predict and not just an empirical fit. A previous-token head performs a position-defined function and so must have $\rho > 0$ (Proposition 1), while a duplicate-token head performs a resemblance test and so must build a symmetric, positive form. The data confirms these predictions. This matches what the two invariants measure. Duplicate and induction heads act on content (low positional capacity), and the two are told apart by α. The previous-token heads are the positional ones and are picked out by ρ alone.

The logistic regression on (ρ, α) separates every pair of head types at ROC-AUC ≥ 0.97 across all five models, with most pairs at 1.00, as seen in Table 1. The lowest values (0.977 for induction vs. duplicate on Qwen2.5-7B and Llama-3.1-8B) are also the pairs closest in the plane. Since a two-feature linear model has little capacity to overfit, this reflects the separability of the coordinates themselves. The two invariants are found to be needed for an accurate classification. Ranked by ρ alone, induction and duplicate heads are not separable (AUC 0.20 to 0.71 across models). Likewise, induction and previous-token heads are not separable when ranked by α alone (AUC 0.33 to 0.96), since that split is positional and α does not make use of positional encoding. Using both coordinates, every pair is separated at AUC ≥ 0.97.

Additionally, we run k-means with $k = 3$ on (ρ, α) and measure how purely the resulting clusters recover the labeled types, as evidenced in Figure 1. Purity ranges from 0.86 to 1.00 across the five models: 1.00 on Qwen2.5-1.5B, 0.96 on Mistral-7B, 0.93 on Gemma-2-2B, 0.91 on Qwen2.5-7B and 0.86 on Llama-3.1-8B. We choose $k = 3$ to match the number of labeled types and report label purity. Even without any labels, the clusters align with the behavioral types, which shows the structure is intrinsic to the invariants and not an artifact of the supervised fit.

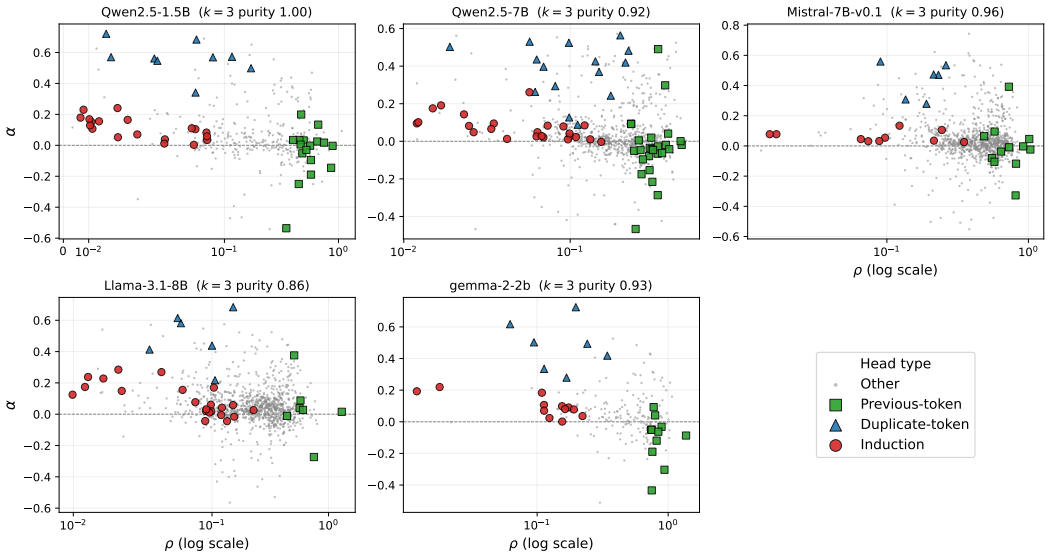


Figure 1: Attention heads in the (ρ, α) plane across five models from four architecture families. Behaviourally labelled heads occupy the same distinct regions in every model: induction (red circles) at low ρ with small positive α, duplicate-token (blue triangles) at low ρ with high α, and previous-token (green squares) at high ρ. Unlabeled heads (grey) fill the remainder of the chart. Panel titles report the unsupervised $k = 3$ label purity. The consistency of this layout across architectures is what the transfer experiment (Section 5.3) quantifies.

Table 1: Pairwise separability ROC-AUC using ρ only, α only, and both, for the three type-pairs induction/duplicate (i/d), induction/previous (i/p), duplicate/previous (d/p). ρ alone is a weak classifier of i/d, α alone is a weak classifier of i/p, and when both invariants are used, every pair is separated.

	ρ only			α only			$\rho + \alpha$		
Model	i/d	i/p	d/p	i/d	i/p	d/p	i/d	i/p	d/p
Qwen2.5-1.5B	0.71	1.00	1.00	1.00	0.86	1.00	1.00	1.00	1.00
Qwen2.5-7B	0.71	1.00	1.00	0.95	0.84	0.94	0.98	1.00	1.00
Mistral-7B	0.60	1.00	1.00	1.00	0.77	0.97	1.00	1.00	1.00
Llama-3.1-8B	0.20	1.00	1.00	0.96	0.33	0.93	0.98	1.00	1.00
Gemma-2-2B	0.58	1.00	1.00	1.00	0.91	1.00	1.00	1.00	1.00

Table 2: The low-ρ region is forbidden to previous-token heads. For each model: the minimum ρ over labeled previous-token heads (their ρ floor), the global minimum ρ over all heads, and the fraction of all heads lying below the previous-token floor. No previous-token head lies in the low-ρ region, which nonetheless contains the majority of the model.

Model	prev-token $\rho_{\min}$	all-head $\rho_{\min}$	% heads below prev-token $\rho_{\min}$
Qwen2.5-1.5B	0.349	0.004	72%
Qwen2.5-7B	0.272	0.014	63%
Mistral-7B	0.487	0.024	54%
Llama-3.1-8B	0.440	0.010	88%
Gemma-2-2B	0.742	0.022	79%

Table 3: Cross-model transfer of the logistic-regression $(\rho, \alpha) \to$ type classifier: trained on the labelled heads of the row model, evaluated on the labelled heads of the column model. Values are balanced accuracy (mean per-class recall; chance $= 1/3$). Diagonal (grey) is within-model 5-fold reference.

Train \ Test	Qwen-1.5B	Qwen-7B	Mistral-7B	Llama-3.1-8B	Gemma-2-2B
Qwen-1.5B	0.96	0.86	0.93	0.94	1.00
Qwen-7B	0.98	0.91	0.90	0.96	0.94
Mistral-7B	0.96	0.69	0.97	0.87	1.00
Llama-3.1-8B	0.98	0.78	0.97	0.87	1.00
Gemma-2-2B	0.94	0.64	0.97	0.90	1.00

We are careful about the scope of this claim. The results above concern how the labeled types sit relative to *one another*; they do not assert that a region of the plane contains *only* heads of a given type. Most heads carry no clear behavioral label and form a continuum (the grey points in Figure 1) that overlaps the labeled regions. This is expected under the capacity reading: a head may have, for instance, the positional geometry of a previous-token head (high ρ) without exhibiting that behavior on our probe, in which case it is unlabeled yet lies in the same region. The invariants thus characterize the geometry a behavior requires, not a guarantee that the behavior is present.

5.2 The low-ρ region is forbidden to positional heads

Proposition 1 predicts a forbidden region. A head whose function requires positional encoding cannot have ρ at or near zero. Table 2 tests this against the labeled previous-token heads. In every model the previous-token heads sit above a strictly positive ρ floor (minimum ρ between 0.27 and 0.74), while a majority of all heads (54% to 88%) lie below that floor. Heads with ρ near zero exist in abundance in every model (global minimum $\rho \leq 0.024$), and no previous-token head is among them. The positional function is absent from exactly the low ρ region in which, by Proposition 1, it cannot be represented.

5.3 Transfer across architectures

Given that our method is completely invariant to the basis in which the weights are represented, we may wonder whether a classifier trained on one model should label heads in another without retraining. Table 3 tests this for all twenty ordered pairs. We fit the logistic regression on the labeled heads of one model and evaluate it on another. The median balanced accuracy is found to be 0.94 and 16 of the 20 pairs reach 0.87 or more, against a chance level of $1/3$, as seen in Table 3. This transfer holds both across architecture families and across scales within one family, thus pointing to (ρ, α) being an architecture-independent coordinate system. The four weakest transfers all have Qwen2.5-7B as the target. In this model, both its own separability and transfer results are the lowest among the 7–8B models. Transfers from Qwen2.5-7B to other models remain high. This transfer also addresses the small number of samples by allowing us to assess these invariants with a more global perspective.

Table 4: Context extension at $S{=}4\times$. Targeted applies YaRN only to the top-50% ρ heads. On both models the targeted variant improves on uniform YaRN at both the short and long context.

Model	Method	Short PPL (512)	Long PPL (4096)
Qwen2.5-1.5B	Baseline ($S{=}1$)	31.66	9.73
	YaRN (all heads)	32.86	11.72
	YaRN + Targeted (ours)	**32.27**	**11.59**
Mistral-7B	Baseline ($S{=}1$)	7.63	5.44
	YaRN (all heads)	8.60	6.83
	YaRN + Targeted (ours)	**8.31**	**6.36**

6 Application: improving YaRN

This section serves as a proof of concept that the invariants are actionable, and is not central to the results of this paper. Using ρ alone allows us to measure how much a head's geometry couples to position, yielding a data-independent way to decide which heads a context-extension method should act on. YaRN (Peng et al., 2024) extends context by rescaling RoPE frequencies uniformly across all heads, which also perturbs heads that do little positional work and can cost short-context quality. We instead apply YaRN only to the top 50% of heads by ρ, leaving the rest at their native frequencies.

Table 4 reports perplexity on Qwen2.5-1.5B and Mistral-7B at a $4\times$ extension. On both models the ρ-targeted variant improves on uniform YaRN at both the short and the long context. On Qwen2.5-1.5B, it recovers about half of the short-context degradation that uniform YaRN introduces (short 32.27 vs. 32.86, long 11.59 vs. 11.72). On Mistral-7B, it shows the same pattern (short 8.31 vs. 8.60, long 6.36 vs. 6.83).

The gain is modest, and it is largest where uniform YaRN hurts short-context quality more, which is exactly where low ρ heads can be spared.

7 Limitations

We note three limitations, stated plainly.

Scope of head types and architectures. We study three head types: induction, duplicate-token, previous-token. We choose them because they are detectable by our invariants and are shown to be model-agnostic. The OV-sector is left out of scope by construction throughout this paper. Both ρ and α are defined through RoPE and thus apply to RoPE architectures only. Extending the symmetry analysis to other positional schemes is left as future work.

Small labeled classes, and scale. Named functional heads are intrinsically rare: only a small fraction of a model's heads cleanly implement a canonical role such as induction or duplicate-token, so any study of these types works with modest counts by nature rather than by a limitation of our labeling. The behaviorally labeled heads are correspondingly small in number for most models, so no single per-model number is decisive. The evidence is shown through the consistency of ρ and α across four architecture families and a $4.7\times$ scale range (Figure 1), together with cross-model transfer (Section 5.3). We emphasize that unsupervised label purity does not degrade monotonically with model size, as within the 7–8B band it ranges from 0.86 (Llama-3.1-8B) to 0.96 (Mistral-7B). That said, our largest model is 8B and because ρ, α are per-head weight functions with no forward pass, they are defined identically at any scale and the cost of computing them grows only with head count. We expect the chart to persist in much larger models, but models at the 70B$+$ scale remain untested here, and confirming the chart is a natural extension.

A single application. We demonstrate one downstream use of the invariants, ρ-targeted context extension, Section 6. An α-based application is left to future work.

8 CONCLUSION

We characterize attention heads with two scalars, ρ and α, which are computed from the query and key weights alone, with no data and no forward pass. These were built from the symmetries of the QK sector: ρ measures how much a head couples to position, and α measures the definiteness of its content interaction, from a similarity kernel ($\alpha \to +1$) to an anti-similarity kernel ($\alpha \to -1$).

On five models from four architecture families, the pair ρ, α places the induction, duplicate-token and previous-token heads in the same three regions of a two-dimensional plane. It successfully separates each pair of types at ROC-AUC ≥ 0.97, and transfers across architectures at a median balanced accuracy of 0.94.

A central result is a necessity law of RoPE attention, valid for most modern LLMs. We prove that $\rho = 0$ leaves the attention logit independent of relative position, and as such, positional function is unrepresentable. Thus, $\rho > 0$ is a necessary condition for positional function (Proposition 1). Furthermore, we demonstrate that across all five models no previous-token head lies in the resulting low ρ region, which nonetheless holds the majority of heads. Finally, as a first use of the invariants, targeting YaRN by ρ improves on uniform YaRN. These are structural, weight-space measurements of what a head is built to do, and they can complement activation-based interpretability.

By working with weight invariants alone we require no data, no prompts and no forward pass, so every head in a model is charted in one cheap pass. Furthermore, the same coordinates apply to any RoPE model without per-model tuning. This makes ρ and α a natural coordinate system for data-free head classification, and opens the possibilities into studying more mappings between capacity and usage of a head.

REPRODUCIBILITY STATEMENT

The invariants ρ and α are deterministic functions of public model weights, defined in Section 3.2 and Section 3.3 and computed as described in Section 4. We release the code, the per-model result files containing ρ, α and the behavioral labels for every head, and the scripts that produce every table and figure in the paper.

USE OF AI ASSISTANCE

We used AI assistance in three ways. First, to aid and polish writing. Second, for retrieval and discovery, for example to help locate related work. Third, to assist in producing the experiment code, which we reviewed and tested. All research ideas, the derivations, the experimental design, and the analysis are our own, and we verified all AI-assisted output, including the code and any referenced prior work.

REFERENCES

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 6(16318-16352):1, 2025.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Yiting Chen and Junchi Yan. What rotary position embedding can tell us: Identifying query and key weights corresponding to basic syntactic or high-level semantic information. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:54507–54528, 2024.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Amit Elhelo and Mor Geva. Inferring functionality of attention heads from their parameters. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17701–17733, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.866. URL https://aclanthology.org/2025.acl-long.866/.

Javier Ferrando, Gerard I Gállego, Ioannis Tsiamas, and Marta R Costa-Jussà. Explaining how transformers use context to build predictions. *arXiv preprint arXiv:2305.12535*, 2023.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Anand Gopalakrishnan, Róbert Csordás, Jürgen Schmidhuber, and Michael C. Mozer. Decoupling the "what" and "where" with polar coordinate positional embeddings. *arXiv preprint arXiv:2509.10534*, 2026. ICML 2026.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. In *International Conference on Learning Representations (ICLR)*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*, 2023.

Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. Base of RoPE bounds context length. *arXiv preprint arXiv:2405.14591*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Mansi Sakarvadia, Arham Khan, Aswathy Ajith, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. Attention lens: A tool for mechanistically interpreting the attention head information retrieval mechanism. *arXiv preprint arXiv:2310.16270*, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Felipe Urrutia, Juan José Alegría, Cinthia Sanchez Macias, Jorge Salas, Cristian B. Calderon, and Cristobal Rojas. Positional versus symbolic attention heads: Learning dynamics, rope geometry, and length generalization. *arXiv preprint arXiv:2605.31558*, 2026a.

Felipe Urrutia, Jorge Salas, Alexander Kozachinskiy, Cristian Buc Calderon, Hector Pasten, and Cristobal Rojas. Decoupling positional and symbolic attention behavior in transformers. In *International Conference on Learning Representations (ICLR)*, 2026b.

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

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.

A THE RESIDUAL SYMMETRY GROUP UNDER ROPE

From equation 2, an orthogonal $\boldsymbol{S} \in O(d_h)$ fixes $\boldsymbol{G}_{\mathrm{QK}}(\Delta)$ for all Δ iff it commutes with every $\boldsymbol{R}(\Delta)$. Since $\boldsymbol{R}(\Delta)$ is a block-diagonal matrix with a distinct rotation frequency ω_k per plane, an orthogonal matrix that commutes with all of them cannot mix planes. Thus, within each plane, $\boldsymbol{S}$ must commute with every 2×2 rotation, and the transformations that do so are exactly the rotations $SO(2)$ (a reflection $\boldsymbol{F}$ conjugates a rotation to its inverse, $\boldsymbol{F}\boldsymbol{R}(\Delta)\boldsymbol{F}^\top = \boldsymbol{R}(-\Delta) \neq \boldsymbol{R}(\Delta)$, and is therefore excluded). As such, the symmetry is broken by RoPE from $O(d_h)$ to $SO(2)^{d_h/2}$.

B DEGREES OF FREEDOM OF THE QK SECTOR

We count the basis-independent quantities available in the QK sector of a single head and locate ρ and α among them.

The QK sector is parameterized by $(\boldsymbol{W}_Q, \boldsymbol{W}_K) \in \mathbb{R}^{d_h \times d} \times \mathbb{R}^{d_h \times d}$, a space of dimension $2d_h d$. By equation 2, the group $G = SO(2)^{d_h/2}$ acts by $(\boldsymbol{W}_Q, \boldsymbol{W}_K) \mapsto (\boldsymbol{S}\boldsymbol{W}_Q, \boldsymbol{S}\boldsymbol{W}_K)$ and leaves $\boldsymbol{G}_{\mathrm{QK}}(\Delta)$ fixed for all Δ. Its dimension is $\dim G = \sum_{k=1}^{d_h/2} \dim SO(2) = d_h/2$, since each $\dim SO(2) = 1$. For $(\boldsymbol{W}_Q, \boldsymbol{W}_K)$ of full rank the continuous stabilizer is trivial, so the action is generically free and the space of inequivalent heads has dimension

$$\dim\left(\{(\boldsymbol{W}_Q, \boldsymbol{W}_K)\}/G\right) \;=\; 2d_h d - \tfrac{d_h}{2}\,. \tag{9}$$

The composite $\boldsymbol{G}_{\mathrm{QK}}(\Delta)$ is already invariant under G, so ρ and α are scalar functions of it. They split along the decomposition $\boldsymbol{G}_{\mathrm{QK}}(\Delta) = \boldsymbol{G}_{\mathrm{QK}}(0) + \boldsymbol{P}(\Delta)$. The positional part factorizes over the $d_h/2$ planes, and $\rho = \sum_k \rho_k$ is the corresponding sum and the static term $\boldsymbol{G}_{\mathrm{QK}}(0)$ has no $\boldsymbol{R}(\Delta)$ dependence. α is a scalar built from the eigenvalues of its symmetric part. Thus ρ and α are the leading scalar invariants of the positional and static parts of $\boldsymbol{G}_{\mathrm{QK}}$ respectively.

C THE SOFTMAX SHIFT-SYMMETRY

Attention has a second exact symmetry that does not act on the weights and does not relate simply to the QK structure. For completeness, we discuss it in this appendix and explain why our invariants

Table 5: Label-threshold robustness. Pairwise ROC-AUC of (ρ, α) separating each head-type pair: induction/duplicate (i/d), induction/previous (i/p), duplicate/previous (d/p). Additionally, we present unsupervised k=3 label purity, per model, across labeling thresholds.

Model	thr	AUC i/d	AUC i/p	AUC d/p	k=3 purity
Qwen2.5-1.5B	0.20	1.00	1.00	1.00	0.99
	0.25	1.00	1.00	1.00	1.00
	0.30	1.00	1.00	1.00	1.00
	0.35	1.00	1.00	1.00	1.00
Qwen2.5-7B	0.20	0.98	0.96	1.00	0.91
	0.25	0.97	1.00	1.00	0.92
	0.30	0.98	1.00	1.00	0.91
	0.35	0.98	1.00	1.00	0.93
Mistral-7B	0.20	1.00	1.00	1.00	0.93
	0.25	1.00	1.00	1.00	0.95
	0.30	1.00	1.00	1.00	0.96
	0.35	1.00	0.97	1.00	0.96
Llama-3.1-8B	0.20	0.99	1.00	0.94	0.86
	0.25	1.00	1.00	1.00	0.83
	0.30	0.98	1.00	1.00	0.86
	0.35	1.00	–	–	0.85
Gemma-2-2B	0.20	1.00	1.00	0.98	0.91
	0.25	1.00	1.00	1.00	0.94
	0.30	1.00	1.00	1.00	0.93
	0.35	1.00	1.00	1.00	0.92

are unaffected in practice. We apply softmax to the attention weights over the key index j of each row $s_{i\cdot}$. Thus, adding any constant c_i leaves them unchanged,

$$s_{ij} \;\longmapsto\; s_{ij} + c_i \qquad \Longrightarrow \qquad \mathrm{softmax}_j(s_{i\cdot} + c_i) = \mathrm{softmax}_j(s_{i\cdot})\,. \tag{10}$$

This shift symmetry acts on the data dependent pre-softmax matrix, and as such, it does not act on $\boldsymbol{G}_{\mathrm{QK}}$ directly and it does not induce a reparameterization of the weights that our invariants must respect. We compute ρ and α directly from $\boldsymbol{G}_{\mathrm{QK}}$ and do not quotient by this shift. A weight-space treatment of this symmetry is left to future work.

D Robustness to the labeling threshold

Head labels have been chosen under the rule label $= \arg\max$ over the three behavioral scores if that score exceeds a threshold, else "other" (main text: 0.30). Table 5 re-labels at thresholds 0.20–0.35 and recomputes pairwise separability and unsupervised k=3 purity. Both stay high throughout (AUC ≥ 0.96; purity ≥ 0.83), demonstrating that the taxonomy is not an artifact of the 0.30 choice. At 0.35, the smallest classes drop below the five-head minimum for a stable 5-fold AUC on Llama-3.1-8B (marked "–").

E Labeled-head counts

Named functional heads are intrinsically rare, so the behaviorally labeled classes are small. Table 6 reports, for each model, the number of heads assigned to each type at the 0.30 threshold used in the main text, together with the total head count. These are the class sizes underlying the separability and transfer results (Tables 1 and 3). We report them so the reported AUC and balanced-accuracy values can be read against the counts they are computed from.

Table 6: Labeled-head counts per model at the 0.30 threshold. Number of heads assigned to induction, duplicate-token, and previous-token, and the model's total head count. Remaining heads are unlabeled ("other").

Model	Induction	Duplicate	Previous-token	Total heads
Qwen2.5-1.5B	19	9	17	336
Qwen2.5-7B	24	16	27	784
Mistral-7B	10	6	12	1024
Llama-3.1-8B	23	6	7	1024
Gemma-2-2B	12	7	11	208