MODERN TRANSFORMERS ARE IMPLICIT HYBRIDS: FROM FUNCTIONAL DIFFERENTIATION TO PRINCIPLED HYBRID ARCHITECTURE DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid architectures combining Full Attention (FA) and Linear Attention (LA) are increasingly prominent, yet their allocation remains largely heuristic. We seek an evidence-grounded basis in the head-level functional organization learned by modern RoPE-based Transformers. Behavioral retrieval and local-streaming probes reveal useful tendencies but do not yield a complete taxonomy. We therefore propose two faithful intervention-based metrics: *RoPE Frequency Importance Score* (RFIS), which measures how each frequency contribution affects a head's complete attention distribution, and *RoPE Positional Dependence* (RPD), which isolates the effect of rotary positional modulation. Applied to Qwen3-series models and Llama3.1, RFIS suggests and RPD verifies a complete two-type taxonomy comprising retrieval and positional heads in RoPE Transformers, separated by a salient mid-low-frequency band. Controlled Transformers show that this functionally separating band follows the training-length positional scale; we term this mechanism-level boundary the *Global Positional Band* (GPBand). Its global positional dependence suggests a potential cause of zero-shot length-extrapolation failure and, together with the layer-specific head distribution, yields two design principles: *(i)* positional modeling should operate only locally, while global access should be implemented through position-independent retrieval; and *(ii)* retrieval and positional functions should be assigned at head granularity with layer-specific allocation. We instantiate these principles in the *Head-wise Hybrid Architecture* (HwH), using NoPE FA for global retrieval and LA for local positional modeling. With an overall FA-to-LA ratio less than 1:3, HwH retains strong language modeling and commonsense reasoning while improving retrieval and substantially strengthening zero-shot long-context extrapolation over Transformer, LA, and a 1:3 layer-wise hybrid baseline. Ablations validate both principles and the component roles, highlighting principled hybrid architecture design as a promising route toward future foundation models.

1 INTRODUCTION

Full Attention (FA) provides exact token-to-token access but incurs quadratic sequence complexity (Vaswani et al., 2017) and a KV cache that grows with context during autoregressive decoding (Shazeer, 2019). Fixed-state LA offers recurrent efficiency (Katharopoulos et al., 2020) but can struggle with exact long-range retrieval (Arora et al., 2024a; Jelassi et al., 2024). Their combination is attractive, but neither mechanism determines how a hybrid should be organized. Existing systems rely on fixed allocations or empirically selected architectural choices (Lieber et al., 2024; Wang et al., 2025; Team et al., 2025; Dong et al., 2025; Zuo et al., 2025). This leaves the central design question unresolved: which roles should be assigned to FA or LA, at what granularity, and with what allocation strategy?

The functional differentiation learned by modern RoPE-based Transformers offers a natural design reference (Figure 1). Prior work identifies retrieval heads that support long-range access (Wu et al., 2025; Zhang et al., 2025b); subsequent methods exploit this specialization for post-training KV-cache optimization or checkpoint conversion (Xiao et al., 2025; Tan et al., 2026). However, these

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.

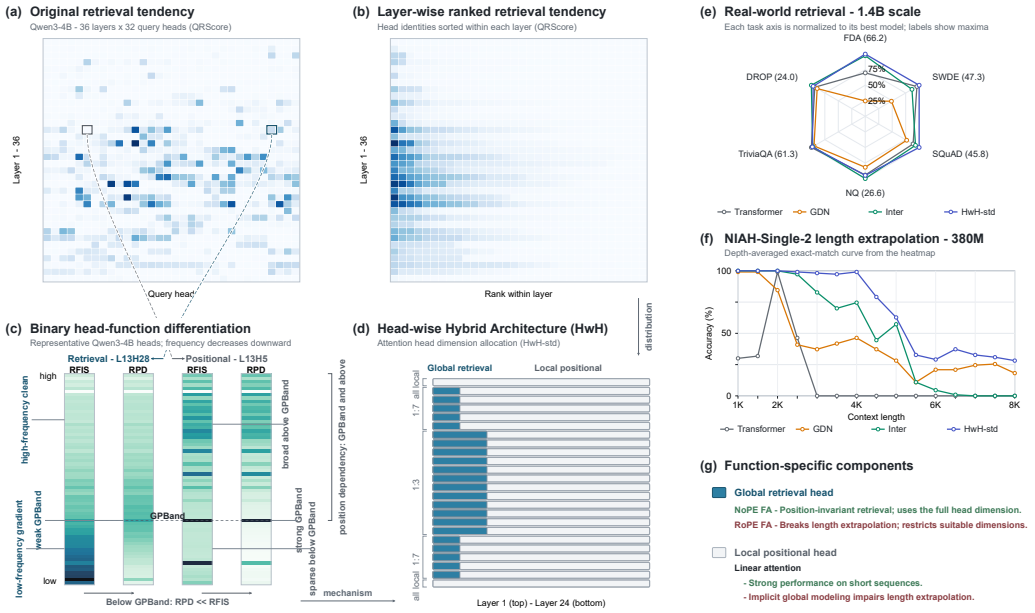

Figure 1: **Overview: from Transformer functional differentiation to principled hybrid design.** **(a)** Retrieval tendency of Qwen3-4B measured by QRscore. **(b)** Layer-specific retrieval distribution after within-layer head sorting. **(c)** RFIS/RPD reveal GPBand-separated retrieval and positional functions. **(d)** HwH instantiates two principles: *(i)* positional modeling should operate only locally, while global access should be implemented through position-independent retrieval; and *(ii)* retrieval and positional functions should be assigned at head granularity with layer-specific allocation. **(e)** Comparison on real-world retrieval benchmarks. **(f)** Zero-shot length extrapolation on NIAH-Single-2. **(g)** NoPE FA serves global retrieval; LA serves local positional modeling.

methods begin with pretrained Transformers and do not independently characterize the complementary function; they therefore cannot by themselves ground architecture design from scratch.

To establish a complete mechanism-level taxonomy, we begin with behavioral probes of head-level functional differentiation. We use QRscore (Zhang et al., 2025b) to measure retrieval tendency and introduce LDscore to measure local-streaming tendency. The two probes do not yield a complete taxonomy, so behavioral evidence alone cannot classify all heads mechanistically. We consequently turn to the underlying RoPE-frequency dependence and propose two faithful intervention-based metrics. *RoPE Frequency Importance Score* (RFIS) measures how removing one frequency contribution changes a head's complete attention distribution, whereas *RoPE Positional Dependence* (RPD) isolates whether that effect arises from rotary positional modulation. RFIS motivates, and RPD directly verifies, a complete two-type taxonomy comprising retrieval and positional heads in RoPE Transformers: the former emphasize frequencies below a salient mid-low-frequency band, whereas the latter emphasize the band and the frequencies above it. Together, these results establish retrieval and positional modeling as complementary mechanism-level functions.

Controlled Transformers further show that this functional boundary follows the training-length positional scale. We therefore term it the *Global Positional Band* (GPBand). The associated global positional fitting suggests a potential failure mode beyond the training length. Together with the depth-varying head distribution in Figure 1b, this yields two principles: *(i)* positional modeling should operate only locally, while global access should be implemented through position-independent retrieval; and *(ii)* retrieval and positional functions should be assigned at head granularity with layer-specific allocation.

We instantiate these principles in the Head-wise Hybrid Architecture (HwH), which assigns NoPE FA to global retrieval and LA to local positional modeling with a layer-specific FA-to-LA ratio no greater than 1:3. From-scratch pretraining shows that HwH retains strong language modeling and

commonsense reasoning, improves retrieval, and substantially strengthens zero-shot extrapolation over Transformer, pure-LA, and layer-wise hybrid baselines. Component and allocation ablations further validate both principles and the roles assigned to FA and LA (Figure 1d–g).

Our contributions are summarized as follows:

- We propose RFIS and RPD, faithful and bounded intervention-based metrics for RoPE-frequency analysis, and use them to establish a complete head-level taxonomy of retrieval and positional modeling in RoPE Transformers.
- We characterize the training-length-related salient mid-low-frequency band in RoPE Transformers as the functional boundary of this taxonomy, term it GPBand, and identify global positional fitting as a potential cause of extrapolation failure.
- We derive two hybrid-design principles and validate them through from-scratch HwH experiments, which also confirm LA's role in local positional modeling and NoPE FA's role in position-independent global retrieval.
- We reveal RoPE Transformers as implicit functional hybrids and motivate principled hybridization with better function-specific components as a promising route toward future foundation models.

2 PRELIMINARIES

Full attention and grouped-query attention. For a sequence $X \in \mathbb{R}^{T\times d}$, FA (Vaswani et al., 2017) forms $Q = XW^Q$, $K = XW^K$, and $V = XW^V$. A head of dimension d_h outputs

$$O = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + M\right) V, \tag{1}$$

where M is the causal mask. FA requires $O(T^2)$ pairwise interactions, and its KV cache grows linearly with context. Grouped-query attention (GQA) (Ainslie et al., 2023) shares each of $G \le H$ KV heads among a group of H Query heads; let $g(h)$ map Query head h to its KV head. The cases $G = H$ and $G = 1$ recover MHA and MQA (Shazeer, 2019), respectively. We measure each Query head separately because it retains its own query projection and attention distribution.

Linear attention. LA replaces pairwise FA with a fixed-size recurrent state (Katharopoulos et al., 2020). A basic recurrence is

$$S_t = S_{t-1} + v_t k_t^\top, \qquad o_t = S_t q_t, \tag{2}$$

where $S_t \in \mathbb{R}^{d_v\times d_k}$, giving linear sequence complexity and constant-size recurrent memory for fixed dimensions. We use Gated DeltaNet (GDN) (Yang et al., 2025b), which adds data-dependent decay and a delta-rule update:

$$S_t = S_{t-1}\big(\alpha_t(I - \beta_t k_t k_t^\top)\big) + \beta_t v_t k_t^\top, \qquad o_t = S_t q_t, \tag{3}$$

Here $\alpha_t, \beta_t \in (0, 1)$ control retention and updating.

Rotary position embedding. Rotary Position Embedding (RoPE) partitions each Query and Key head into $F = d_h/2$ two-dimensional groups and rotates group $r \in \{0, \dots, F-1\}$ by frequency ω_r (Su et al., 2023):

$$\begin{aligned} \omega_r &= \theta^{-2r/d_h}, \qquad R(\varphi) = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix}, \\ \tilde{q}_t^{(r)} &= R(t\omega_r) q_t^{(r)}, \qquad \tilde{k}_s^{(r)} = R(s\omega_r) k_s^{(r)}. \end{aligned} \tag{4}$$

Here θ is the RoPE base and $r = 0$ is the highest frequency. The resulting dot product decomposes into exact frequency contributions,

$$\tilde{q}_t^\top \tilde{k}_s = \sum_{r=0}^{F-1} \left(q_t^{(r)}\right)^\top R((s-t)\omega_r)\, k_s^{(r)}, \tag{5}$$

so each group has rotation period $2\pi/\omega_r$ and a distinct positional scale.

3 FUNCTIONAL DIFFERENTIATION AND PRINCIPLES FOR HYBRID DESIGN

3.1 BEHAVIORAL PROBES OF ATTENTION-HEAD TENDENCIES

We measure each Query head's retrieval tendency with QRscore, based on query-to-evidence attention (Zhang et al., 2025b), and its local-streaming tendency with LDscore, based on the fraction of non-sink attention within a recent window (Appendix A.1).

Qwen3-4B is the primary model for analysis; Qwen3-1.7B, Qwen3-8B, and Llama3.1-8B-Instruct provide cross-scale and cross-family checks (Appendix A.3). Its retrieval tendency is sparse, absent from the boundary layers, and concentrated near the middle (Figure 1a–b). Because retrieval and local-streaming tendencies are not globally complementary (Appendix A.2), they provide behavioral anchors and a layer-specific allocation pattern, but not a complete taxonomy.

3.2 RFIS REVEALS RETRIEVAL–POSITIONAL FUNCTIONAL DIFFERENTIATION

To characterize the RoPE-frequency dependence underlying these behavioral tendencies, we introduce the *RoPE Frequency Importance Score* (RFIS). RFIS removes one two-dimensional frequency contribution from a head's logits while holding every other contribution fixed, then measures the resulting change in the complete attention distribution. Formally, for sample x, let $\mathcal{C}_x(i)$ be the keys visible to query position i. At layer ℓ, Query head h, and frequency r, with $g(h)$ its GQA KV head, the exact contribution to the query–key logit is

$$c^{\mathrm{RoPE}}_{\ell,h,i,j,r} = \mathbf{q}^{\top}_{\ell,h,i,r} R((j-i)\omega_r)\, \mathbf{k}_{\ell,g(h),j,r}. \tag{6}$$

The complete logits and attention distribution over the visible keys are

$$z_{\ell,h,i,j} = \sum_{r=0}^{F-1} c^{\mathrm{RoPE}}_{\ell,h,i,j,r}, \qquad \mathbf{p}_{\ell,h,i} = \mathrm{softmax}(\mathbf{z}_{\ell,h,i}). \tag{7}$$

Removing frequency r gives

$$\mathbf{z}^{(-r)}_{\ell,h,i} = \mathbf{z}_{\ell,h,i} - \mathbf{c}^{\mathrm{RoPE}}_{\ell,h,i,r}, \qquad \mathbf{p}^{(-r)}_{\ell,h,i} = \mathrm{softmax}\Big(\mathbf{z}^{(-r)}_{\ell,h,i}\Big). \tag{8}$$

For distributions $\mathbf{a}, \mathbf{b}$ and $\mathbf{m} = (\mathbf{a}+\mathbf{b})/2$, let

$$D_{\mathrm{JS}}(\mathbf{a},\mathbf{b}) = \frac{1}{2} D_{\mathrm{KL}}(\mathbf{a}\|\mathbf{m}) + \frac{1}{2} D_{\mathrm{KL}}(\mathbf{b}\|\mathbf{m}). \tag{9}$$

The per-query and aggregate scores are

$$I^{\mathrm{RFIS}}_{\ell,h,i,r} = \frac{D_{\mathrm{JS}}\Big(\mathbf{p}_{\ell,h,i}, \mathbf{p}^{(-r)}_{\ell,h,i}\Big)}{\ln 2}, \qquad \mathrm{RFIS}_{\ell,h,r} = \mathbb{E}_{x\sim\mathcal{D}}\, \mathbb{E}_{i\sim\mathcal{Q}(x)} \big[I^{\mathrm{RFIS}}_{\ell,h,i,r} \big]. \tag{10}$$

The nested expectation weights samples equally. RFIS is a faithful intervention-based metric naturally bounded in $[0,1]$; larger values indicate stronger dependence on information represented by that frequency. Appendix C gives further theoretical details.

The RFIS panels in Figure 2 show a salient mid-low-frequency band in both representative layers. Retrieval heads depend more on frequencies below it, with a general tendency toward greater importance at lower frequencies, and show little dependence on the band or above. Non-retrieval heads instead emphasize the band and higher frequencies; LDscore has no comparably clear relation. Since retrieval requires stable, approximately position-invariant semantic matching, these patterns suggest a mechanism-level differentiation between retrieval and positional heads. However, RFIS alone cannot verify that the latter pattern depends specifically on rotary modulation, motivating RPD.

3.3 RPD VERIFIES THE POSITIONAL INTERPRETATION OF RFIS

RFIS identifies important frequency contributions but does not by itself determine whether their effects depend on position. We therefore introduce *RoPE Positional Dependence* (RPD), which replaces the relative rotation of one frequency with the identity while preserving its learned Q/K coordinates:

$$c^{\mathrm{NoPE}}_{\ell,h,i,j,r} = \mathbf{q}^{\top}_{\ell,h,i,r} \mathbf{k}_{\ell,g(h),j,r}. \tag{11}$$

The resulting single-frequency RoPE-to-NoPE intervention is

$$\mathbf{z}^{(r\to\mathrm{NoPE})}_{\ell,h,i} = \mathbf{z}_{\ell,h,i} - \mathbf{c}^{\mathrm{RoPE}}_{\ell,h,i,r} + \mathbf{c}^{\mathrm{NoPE}}_{\ell,h,i,r}, \qquad \mathbf{p}^{(r\to\mathrm{NoPE})}_{\ell,h,i} = \mathrm{softmax}\left(\mathbf{z}^{(r\to\mathrm{NoPE})}_{\ell,h,i}\right). \tag{12}$$

We define

$$I^{\mathrm{RPD}}_{\ell,h,i,r} = \frac{D_{\mathrm{JS}}\left(\mathbf{p}_{\ell,h,i}, \mathbf{p}^{(r\to\mathrm{NoPE})}_{\ell,h,i}\right)}{\ln 2}, \qquad \mathrm{RPD}_{\ell,h,r} = \mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{i\sim\mathcal{Q}(x)}\left[I^{\mathrm{RPD}}_{\ell,h,i,r}\right]. \tag{13}$$

RPD is likewise a faithful intervention-based metric bounded in $[0, 1]$, targeting rotation dependence rather than total channel importance. Measurement settings and further theoretical details are in Appendices B.1 and C.

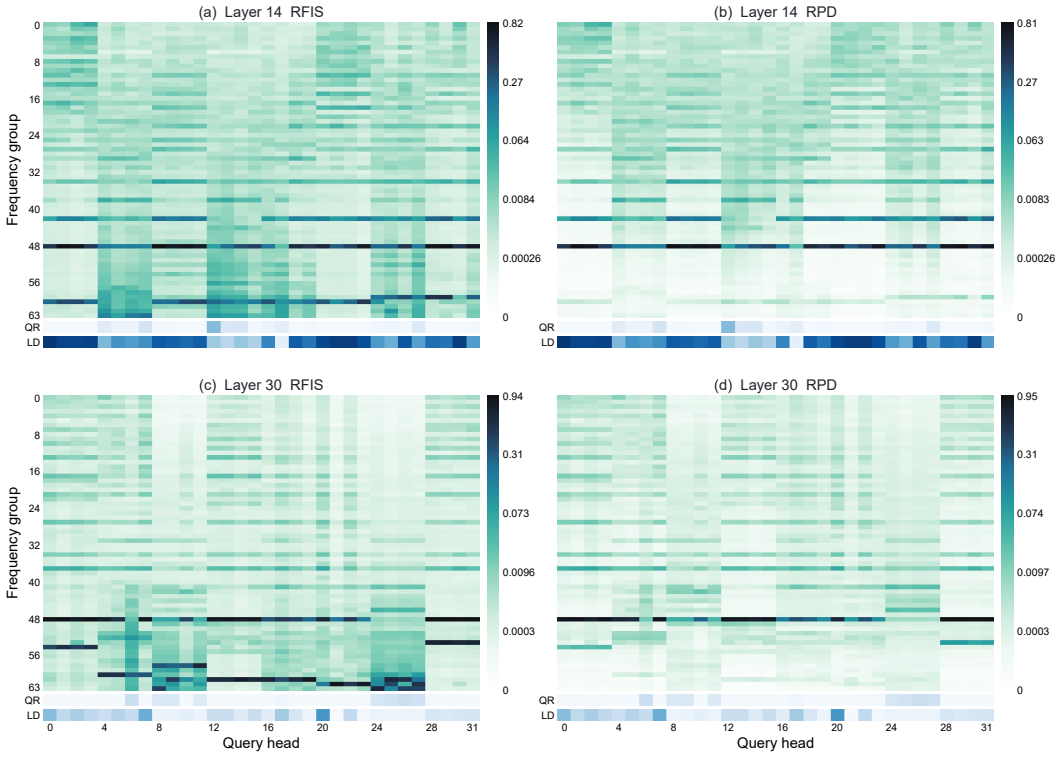

Figure 2: **RFIS and RPD in Qwen3-4B Layers 14 and 30.** RFIS (left) separates below-band retrieval dependence from band-and-above importance; RPD (right) identifies the latter as positional. QRscore/LDscore strips are aligned below each heatmap; frequencies descend from high to low, and panels use independent zero-to-maximum scales.

RPD follows RFIS at the salient band and above, confirming that this pattern depends on rotary positional modulation, while below-band RPD is nearly zero (Figure 2, right). This verifies the RFIS hypothesis: non-retrieval heads with band-and-above dependence are positional, whereas retrieval heads use approximately position-invariant low-frequency information. RFIS and RPD thus establish a complete two-type taxonomy of retrieval and positional heads in RoPE Transformers.

3.4 Global Positional Band

Because the salient band separates positional dependence from approximately position-invariant retrieval dependence, we hypothesize that its positional scale is tied to the context length over which global positional structure is learned. Production LLMs cannot isolate this relation because context-extension training produces multiple depth-varying bands (Appendix B.2). We therefore train two matched 380M-parameter QKNorm Transformers at sequence length 2048, changing only the RoPE base from $\theta = 10^4$ to 10^6.

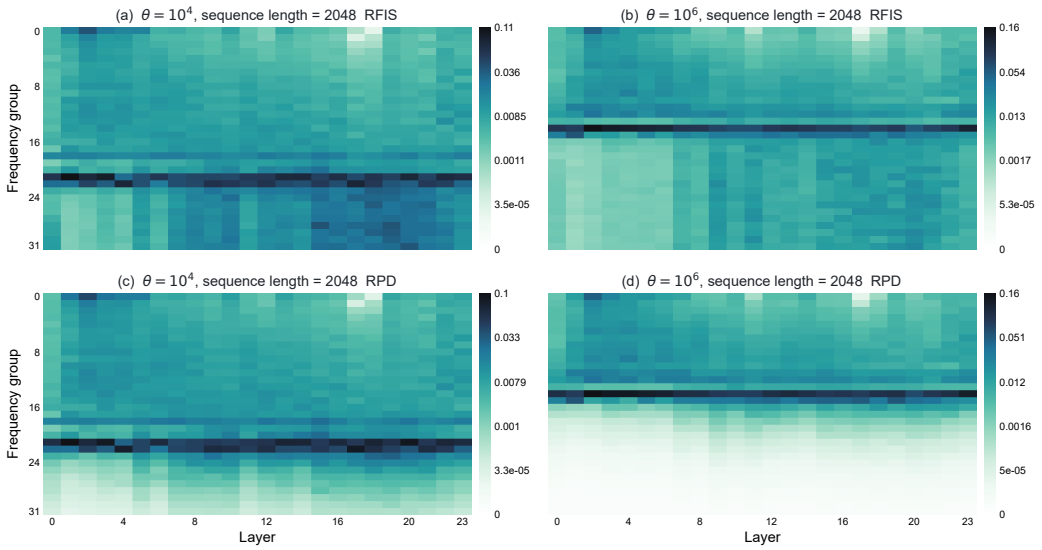

Figure 3: **Controlled validation of the salient band's positional scale.** Layer-mean RFIS (top) and RPD (bottom) in matched Transformers. Changing θ from 10^4 to 10^6 moves the band from groups 21–22 to 14–15 while preserving training-length-scale periods; panels use independent zero-to-maximum scales.

At $\theta = 10^4$, groups 21–22 have periods 2650 and 3533 tokens; at $\theta = 10^6$, groups 14–15 have periods 2650 and 4080. The indices change, but both RFIS/RPD bands remain at the scale of the 2048-token training context. The intervention results therefore establish that this training-length-related band marks the mechanism-level boundary between retrieval and positional dependence. We refer to this functionally characterized boundary as the *Global Positional Band* (GPBand). Derivations and controls for QKNorm and training length appear in Appendix B.4.

Local positional-modeling hypothesis. GPBand shows that positional heads fit global dependencies at the training scale; beyond that range, the learned representation may fail to generalize, making global positional fitting a potential cause of extrapolation failure. We therefore hypothesize that positional modeling need only be local, while global sequence modeling can be completed through position-independent retrieval.

Architectural principles. Together with the coexistence and depth-varying distribution of the two head types, this analysis yields two principles: *(i)* positional modeling should operate only locally, while global access should be implemented through position-independent retrieval; and *(ii)* retrieval and positional functions should be assigned at head granularity with layer-specific allocation. Section 4 tests them through pretraining from scratch.

4 Empirical Validation with the Head-Wise Hybrid Architecture

4.1 Head-Wise Hybrid Architecture

HwH instantiates the two principles from Section 3 as an architectural inductive bias. Within each layer, NoPE FA heads provide position-independent global retrieval and LA (GDN) heads provide local positional modeling; their outputs are concatenated before the standard output projection. These component assignments are tested by the ablations in Section 4.3.

The standard configuration, *HwH-std*, follows the observed layer-specific distribution: the first and last layers use only GDN; the middle half uses NoPE-FA:GDN = 1:3; and the remaining shallow and deep layers use 1:7. Ratios denote allocated Q/K/V dimensions, so FA occupies at most one quarter of a layer. Overall, HwH-std allocates less FA than Inter, the 1:3 layer-wise baseline used in our comparisons.

4.2 MAIN COMPARISON

We pretrain matched 380M and 1.4B models for 15B and 100B tokens, respectively, on FineWeb-Edu (Penedo et al., 2024), using the Mistral tokenizer with a vocabulary size of 32,000 and a 2K context length under the same standard training protocol. We compare HwH-std with a RoPE Transformer, pure GDN, and Inter, which alternates three GDN layers with one NoPE FA layer. Full training and evaluation details appear in Appendix D.

Language modeling. HwH-std matches or exceeds all baselines at 380M and remains competitive at 1.4B (Table 1), retaining strong language modeling and commonsense reasoning across scales.

Table 1: Language modeling perplexity (↓) and commonsense-reasoning accuracy (%, ↑). HellaSwag and ARC-C use normalized accuracy; best and second-best results are bold and underlined.

Model	Wiki ↓	LAMB. ↓	LAMB.	HellaS.	PIQA	ARC-E	ARC-C	WinoG.	OBQA	Avg.
380M parameters; 15B training tokens										
Transformer	29.10	38.26	**33.09**	39.05	66.54	56.90	27.73	**51.22**	21.60	42.30
GDN	28.71	36.03	31.36	39.55	67.03	**58.12**	26.88	49.33	**23.00**	42.18
Inter	27.69	36.05	31.65	39.52	65.83	57.87	27.30	50.28	22.00	42.07
HwH-std	**27.41**	**34.84**	32.76	**39.93**	**67.57**	57.87	**28.16**	50.75	22.40	**42.78**
1.4B parameters; 100B training tokens										
Transformer	17.36	13.31	46.30	55.30	**72.42**	69.87	38.31	56.91	**29.20**	52.61
GDN	17.32	13.42	44.19	55.54	72.20	71.30	36.69	**58.96**	27.60	52.35
Inter	**16.82**	**11.97**	**47.68**	**55.99**	72.36	**71.55**	**38.48**	58.48	26.80	**53.05**
HwH-std	16.85	12.11	**47.68**	55.96	72.09	70.12	36.60	58.25	28.20	52.70

Retrieval. Pure GDN is weakest on real-world retrieval; Inter outperforms Transformer on most tasks, while HwH-std achieves the best overall result (Table 2). On RULER NIAH, Transformer scores zero on all three tasks at 4K and therefore exhibits no zero-shot length extrapolation. HwH-std maintains strong in-distribution retrieval and performs best at 4K, demonstrating both excellent overall retrieval and substantially stronger length extrapolation.

Table 2: Real-world retrieval and RULER NIAH at the 1.4B scale.

	Real-world retrieval							NIAH-Single-1			NIAH-Single-2			NIAH-Single-3		
Model	FDA	SWDE	SQuAD	NQ	TriviaQA	DROP	Avg.	1K	2K	4K	1K	2K	4K	1K	2K	4K
Transformer	46.19	45.27	40.68	25.12	**61.26**	22.47	40.17	**100.0**	**100.0**	0.0	**100.0**	99.8	0.0	70.8	**67.6**	0.0
GDN	16.33	22.95	35.22	21.79	58.23	21.42	29.33	**100.0**	**100.0**	99.4	99.8	85.0	35.2	**87.0**	52.6	23.0
Inter	64.43	40.95	42.73	**26.64**	59.72	**24.01**	43.08	**100.0**	**100.0**	99.8	**100.0**	**100.0**	79.6	39.8	67.4	35.8
HwH-std	**66.15**	**47.25**	**45.84**	25.28	60.13	23.19	**44.64**	**100.0**	**100.0**	**100.0**	99.8	**100.0**	**99.6**	69.2	58.4	**46.2**

Zero-shot length extrapolation. To compare zero-shot length extrapolation across architectures more closely, we evaluate the 380M models with a NIAH-Single-2 heatmap over sequence lengths from 1024 to 8192 and depth percentages from 0% (shallowest, sequence end) to 100% (deepest, sequence beginning), with 10 examples per cell. Figure 4 reports exact-match accuracy.

Because the 1.4B Transformer already exhibits no zero-shot extrapolation on NIAH-Single-2 in Table 2, we omit its detailed 380M heatmap from the main-text comparison. GDN maintains extrapolation above 2K, but only when the target lies within an approximately training-length-sized local window near the sequence end. Inter outperforms GDN through 5K and exhibits a clear lost-in-the-middle pattern (Liu et al., 2024), and its extrapolation nearly vanishes beyond 5K. HwH-std extrapolates nearly perfectly through 4K—twice the training sequence length—and achieves the strongest performance among all models beyond 4K. Together with the language-modeling and retrieval results, these findings provide empirical validation of both principles formulated in Section 3. The following ablations isolate the component roles and layer-specific allocation.

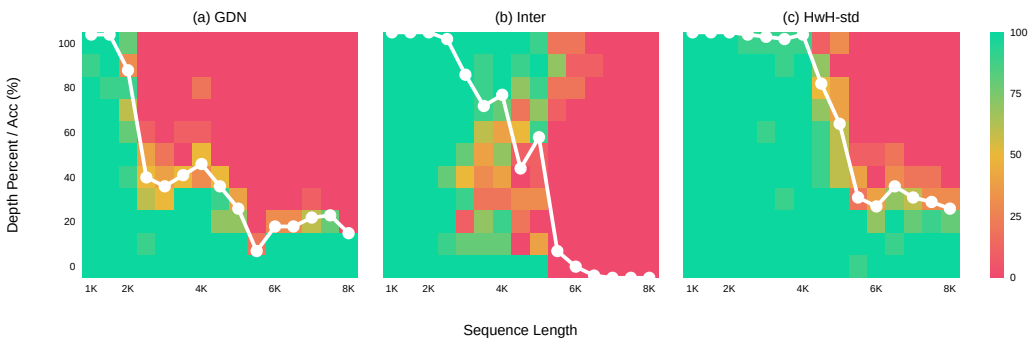

Figure 4: **NIAH-Single-2 accuracy for GDN, Inter, and HwH-std.** Depth runs from 0% (sequence end; bottom) to 100% (sequence beginning; top); white curves are depth averages on the same scale.

4.3 Function-Specific Components

At 380M, *HwH-swa* replaces GDN with sink-free sliding-window attention (SWA; window 128) with RoPE, while *HwH-std-r* replaces NoPE FA with RoPE FA. The first comparison tests LA's role in local positional modeling; the second tests position-independent FA for global retrieval. Figure 5 summarizes task averages and depth-averaged NIAH curves. Detailed task-level results and complete 380M NIAH-Single-2 heatmaps are provided in Appendix E.1.

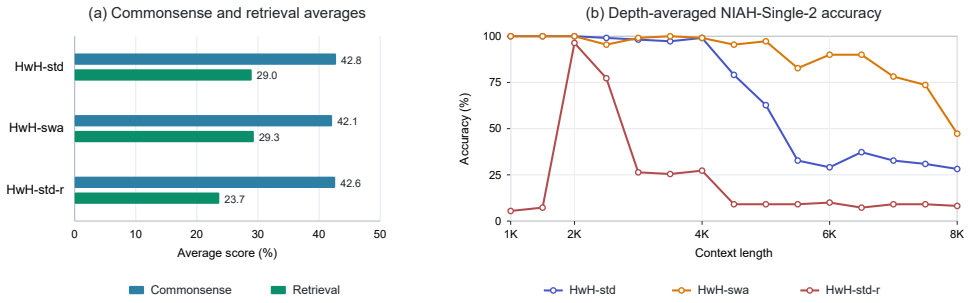

Figure 5: **Function-specific component ablations at** 380**M: (a)** task averages and **(b)** depth-averaged NIAH-Single-2 accuracy.

GDN is stronger than SWA in commonsense reasoning, whereas SWA slightly improves retrieval and substantially improves extrapolation. Thus, GDN contributes effective local positional modeling, while its implicit global state neither aids retrieval nor extrapolation. Replacing NoPE FA with RoPE FA sharply degrades retrieval and extrapolation without improving general modeling ability, confirming NoPE FA's advantage for position-independent global retrieval and the cost of global RoPE positional modeling.

4.4 Layer-Specific Allocation

HwH-uni uses NoPE-FA:GDN $=$ 1:3 in every layer; *HwH-rmfl* removes NoPE FA from the first and last layers; and *HwH-std* additionally uses 1:7 outside the middle half. This progression isolates the layer-specific allocation principle. Figure 6 summarizes the comparison; detailed results are in Appendix E.1.

All variants retain comparable commonsense-reasoning ability. Removing NoPE FA from the first and last layers markedly improves extrapolation, especially through 4K; reducing its outer-layer ratio from 1:3 to 1:7 further improves extrapolation beyond 4K and strengthens retrieval. These results support the layer-specific allocation principle motivated by the depth-varying functional differentiation observed in RoPE Transformers.

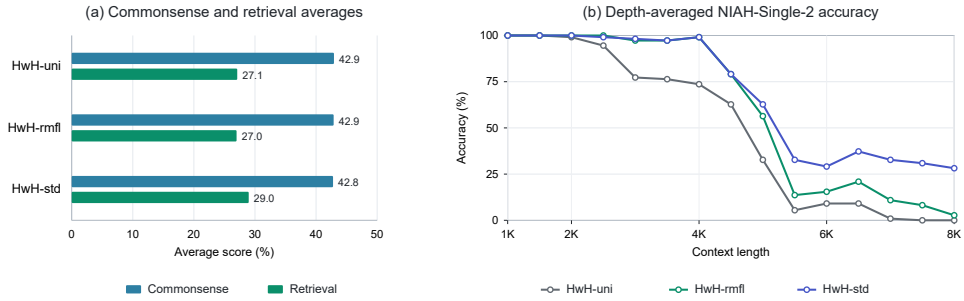

Figure 6: **Layer-specific allocation ablations at** 380**M: (a)** task averages and **(b)** depth-averaged NIAH-Single-2 accuracy.

5 Related Works

Hybrid Architectures. Hybrid architectures differ primarily in allocation granularity. At whole-layer granularity, Lieber et al. (2024), MiniMax et al. (2025), Waleffe et al. (2024), Team et al. (2025), and Qwen Team (2026) interleave FA and LA layers. Wang et al. (2025) systematically analyze layer-wise allocations and recommend LA-to-FA ratios of 3:1–6:1, while Chen et al. (2026) combine inter-layer hybridization with refinements to positional encoding and architecture in Qwen3 conversion. A related softmax-only local–global hybrid, Puvvada et al. (2025), interleaves RoPE SWA with NoPE FA and applies dynamic attention-score scaling at inference. Within layers, Hua et al. (2022), Zancato et al. (2024), Munkhdalai et al. (2024), Nunez et al. (2025), Dong et al. (2025), Zhang et al. (2025a), and Zuo et al. (2025) combine FA and LA through parallel branches or distinct head types. At head granularity, Tan et al. (2026) retain FA specifically at retrieval-critical heads when converting pretrained Transformers. Beyond fixed structural allocation, Li et al. (2026) vary the active mechanism across layers and token positions.

Mechanistic Analysis of Attention Heads and RoPE. Prior work examines attention-head functions and their relation to RoPE. Wu et al. (2025) identify retrieval heads and causally link them to long-context recall; Xiao et al. (2025) detect them with trainable gates and assign other heads to streaming; Zhang et al. (2025b) extend detection to query-to-evidence behavior; and Tan et al. (2026) assess head necessity through activation and path patching. Hong et al. (2024) connect non-uniform RoPE frequency use to token distance. Barbero et al. (2025) use per-frequency Q/K norms as approximate indicators of frequency usage, linking high frequencies to positional behavior and low frequencies to semantic matching. Urrutia et al. (2025) derive per-frequency positional and symbolic scores through input permutation. Both accounts separate low-frequency semantic or symbolic behavior from high-frequency positional behavior, leaving the middle frequencies without a sufficient mechanism-level explanation. Using the same norm-based analysis, Oka et al. (2026) observe the salient band and relate its location theoretically to the RoPE base and training length.

6 Conclusion

We establish a complete head-level taxonomy of retrieval and positional modeling in modern RoPE-based Transformers and identify the training-length-related GPBand as its functional boundary. This analysis yields two hybrid-design principles: *(i)* positional modeling should operate only locally, while global access should be implemented through position-independent retrieval; and *(ii)* retrieval and positional functions should be assigned at head granularity with layer-specific allocation. From-scratch HwH experiments validate these principles and the respective roles of LA and NoPE FA, supporting principled hybrid design for future foundation models.

Future work may extend these analyses and principles to function-aligned GQA, other modalities, larger scales, improved components, and hybrid designs beyond multi-head Transformers.

Reproducibility Statement

We provide the implementation and experiment configurations in the supplementary material, with a corresponding anonymous repository available at `https://anonymous.4open.science/r/Modern-Transformers-Are-Implicit-Hybrids`. Future updates to the repository may include model checkpoints (e.g., Hugging Face paths), detailed QR/LD/RFIS/RPD results and visualizations, and synchronized code revisions. Detailed settings for the behavioral probes, RFIS and RPD analyses, controlled GPBand experiments, model pretraining, evaluation protocols, and HwH variants are documented in the corresponding sections of the Appendix. All architecture comparisons at each model scale use matched training data, optimization schedules, and evaluation protocols to facilitate controlled reproduction of the reported results.

AI Use Statement

Generative AI tools were used to assist with code implementation, language polishing, LaTeX formatting, and figure preparation. The authors reviewed all AI-assisted outputs, tested the AI-assisted code, and take full responsibility for the final manuscript.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. *arXiv preprint arXiv:2402.18668*, 2024a.

Simran Arora, Aman Timalsina, Aaryan Singhal, Benjamin Spector, Sabri Eyuboglu, Xinyi Zhao, Ashish Rao, Atri Rudra, and Christopher Ré. Just Read Twice: Closing the Recall Gap for Recurrent Language Models. *arXiv preprint arXiv:2407.05483*, 2024b.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and Round We Go! What makes Rotary Positional Encodings useful? In *The Thirteenth International Conference on Learning Representations*, 2025.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about Physical Commonsense in Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, 2020.

Yingfa Chen, Zhen Leng Thai, Zihan Zhou, Zhu Zhang, Xingyu Shen, Shuo Wang, Chaojun Xiao, Xu Han, and Zhiyuan Liu. Hybrid Linear Attention Done Right: Efficient Distillation and Effective Architectures for Extremely Long Contexts. *arXiv preprint arXiv:2601.22156*, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Celine Lin, Jan Kautz, and Pavlo Molchanov. Hymba: A Hybrid-head Architecture for Small Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to Train Long-Context Language Models (Effectively). *arXiv preprint arXiv:2410.02660*, 2025.

gkamradt. Gkamradt/LLMTest_NeedleInAHaystack. GitHub repository, 2026. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On the token distance modeling ability of higher RoPE attention dimension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5877–5888, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? *arXiv preprint arXiv:2404.06654*, 2024.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer Quality in Linear Time. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 9099–9117, 2022.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat After Me: Transformers are Better than State Space Models at Copying. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 21502–21521, 2024.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5156–5165, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Yixing Li, Ruobing Xie, Zhen Yang, Xingwu Sun, Shuaipeng Li, Weidong Han, Zhanhui Kang, Yu Cheng, Chengzhong Xu, Di Wang, and Jie Jiang. TransMamba: A Sequence-Level Hybrid Transformer-Mamba Language Model. *arXiv preprint arXiv:2503.24067*, 2026.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A Hybrid Transformer-Mamba Language Model. *arXiv preprint arXiv:2403.19887*, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*, 2017.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, et al. MiniMax-01: Scaling Foundation Models with Lightning Attention. *arXiv preprint arXiv:2501.08313*, 2025.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.

Elvis Nunez, Luca Zancato, Benjamin Bowman, Aditya Golatkar, Wei Xia, and Stefano Soatto. Expansion Span: Combining Fading Memory and Retrieval in Hybrid State Space Models. *arXiv preprint arXiv:2412.13328*, 2025.

Yui Oka, Istumi Saito, Kyosuke Nishida, and Kuniko Saito. Frequency Bands in RoPE: Base Frequency and Context Length Shape the Interpolation–Extrapolation Trade-Off. In *The Fourteenth International Conference on Learning Representations*, 2026.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint arXiv:2406.17557*, 2024.

Krishna C Puvvada, Faisal Ladhak, Santiago Akle Serano, Cheng-Ping Hsieh, Shantanu Acharya, Somshubra Majumdar, Fei Jia, Samuel Kriman, Simeng Sun, Dima Rekesh, and Boris Ginsburg. SWAN: An Efficient and Scalable Approach for Long-Context Language Modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2424–2438, 2025.

Qwen Team. Qwen/Qwen3.6-27B · Hugging Face. Hugging Face model card, 2026. URL `https://huggingface.co/Qwen/Qwen3.6-27B`.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *arXiv preprint arXiv:1907.10641*, 2019.

Noam Shazeer. Fast Transformer Decoding: One Write-Head is All You Need. *arXiv preprint arXiv:1911.02150*, 2019.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv preprint arXiv:2104.09864*, 2023.

Lintang Sutawika, Leo Gao, Hailey Schoelkopf, Stella Biderman, Jonathan Tow, Baber Abbasi, ben fattori, Charles Lovering, farzanehnakhaee70, Jason Phang, Anish Thite, Fazz, Aflah, Niklas Muennighoff, Thomas Wang, sdtblck, nopperl, gakada, tttyuntian, researcher2, Chris, Julen Etxaniz, Zdeněk Kasner, Khalid, Jeffrey Hsu, AndyZwei, Pawan Sasanka Ammanamanchi, Dirk Groeneveld, Ethan Smith, and Eric Tang. EleutherAI/lm-evaluation-harness: Major refactor. Zenodo, 2023. URL `https://zenodo.org/records/10256836`.

Zhentao Tan, Wei Chen, Jingyi Shen, Yao Liu, Xu Shen, Yue Wu, and Jieping Ye. HydraHead: From Head-Level Functional Heterogeneity to Specialized Attention Hybridization. *arXiv preprint arXiv:2606.20097*, 2026.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, et al. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*, 2024.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, et al. Kimi Linear: An Expressive, Efficient Attention Architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Felipe Urrutia, Jorge Salas, Alexander Kozachinskiy, Cristian Buc Calderon, Hector Pasten, and Cristobal Rojas. Decoupling Positional and Symbolic Attention Behavior in Transformers. *arXiv preprint arXiv:2511.11579*, 2025.

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*, volume 30, 2017.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, Garvit Kulshreshtha, Vartika Singh, Jared Casper, Jan Kautz, Mohammad Shoeybi, and Bryan Catanzaro. An Empirical Study of Mamba-based Language Models. *arXiv preprint arXiv:2406.07887*, 2024.

Dustin Wang, Rui-Jie Zhu, Steven Abreu, Yong Shan, Taylor Kergan, Yuqi Pan, Yuhong Chou, Zheng Li, Ge Zhang, Wenhao Huang, and Jason Eshraghian. A Systematic Analysis of Hybrid Linear Attention. *arXiv preprint arXiv:2507.06457*, 2025.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval Head Mechanistically Explains Long-Context Factuality. In *The Thirteenth International Conference on Learning Representations*, 2025.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads. In *The Thirteenth International Conference on Learning Representations*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025a.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. *arXiv preprint arXiv:2412.06464*, 2025b.

Luca Zancato, Arjun Seshadri, Yonatan Dukler, Aditya Golatkar, Yantao Shen, Benjamin Bowman, Matthew Trager, Alessandro Achille, and Stefano Soatto. B'MOJO: Hybrid State Space Realizations of Foundation Models with Eidetic and Fading Memory. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130433–130462, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

Michael Zhang, Simran Arora, Rahul Chalamala, Alan Wu, Benjamin Spector, Aaryan Singhal, Krithik Ramesh, and Christopher Ré. LoLCATs: On Low-Rank Linearizing of Large Language Models. *arXiv preprint arXiv:2410.10254*, 2025a.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-Focused Retrieval Heads Improve Long-Context Reasoning and Re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23791–23805, 2025b.

Jingwei Zuo, Maksim Velikanov, Ilyas Chahed, Younes Belkada, Dhia Eddine Rhayem, Guillaume Kunsch, Hakim Hacid, Hamza Yous, Brahim Farhat, Ibrahim Khadraoui, Mugariya Farooq, Giulia Campesan, Ruxandra Cojocaru, Yasser Djilali, Shi Hu, Iheb Chaabane, Puneesh Khanna, Mohamed El Amine Seddik, Ngoc Dung Huynh, Phuc Le Khac, Leen AlQadi, Billel Mokeddem, Mohamed Chami, Abdalgader Abubaker, Mikhail Lubinets, Kacper Piskorski, and Slim Frikha. Falcon-H1: A Family of Hybrid-Head Language Models Redefining Efficiency and Performance. *arXiv preprint arXiv:2507.22448*, 2025.

A Detailed Behavioral Analysis

A.1 Behavioral Probe Definitions and Measurement Settings

QRscore. We follow the query-focused retrieval-head detection method of Zhang et al. (2025b). Unlike copy-paste-based detection on synthetic Needle-in-a-Haystack tasks (Wu et al., 2025),

QRscore measures retrieval tendency through query-context attention on realistic in-context retrieval tasks.

Consider a prompt $\{\mathcal{D}, q\}$, where $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ is a sequence of candidate passages followed by query q, and let $\mathcal{D}^*(q) \subseteq \mathcal{D}$ be the gold-passage set. For layer ℓ and Query head h, let $A_{\ell,h} \in \mathbb{R}^{T\times T}$ denote the post-softmax causal attention matrix, and let $\mathcal{Q}(q)$ and $\mathcal{D}_i$ denote the token positions of q and d_i, respectively. Following the original definition, the query-focused retrieval score toward passage d_i is

$$\mathrm{QRscore}_{\ell,h}(q, d_i) = \frac{1}{|\mathcal{Q}(q)|} \sum_{u\in\mathcal{Q}(q)} \sum_{v\in\mathcal{D}_i} A_{\ell,h}(u, v). \tag{14}$$

Aggregating over the gold passages gives the score for query q:

$$\mathrm{QRscore}_{\ell,h}(q) = \frac{1}{|\mathcal{Q}(q)|} \sum_{d_i\in\mathcal{D}^*(q)} \sum_{u\in\mathcal{Q}(q)} \sum_{v\in\mathcal{D}_i} A_{\ell,h}(u, v). \tag{15}$$

For an evaluation dataset $\mathcal{T} = \{(q, \mathcal{D}, \mathcal{D}^*(q))\}$, the head-level score is

$$\mathrm{QRscore}_{\ell,h,\mathcal{T}} = \frac{1}{|\mathcal{T}|} \sum_{(q,\mathcal{D},\mathcal{D}^*(q))\in\mathcal{T}} \mathrm{QRscore}_{\ell,h}(q). \tag{16}$$

We abbreviate $\mathrm{QRscore}_{\ell,h,\mathcal{T}}$ as $\mathrm{QRscore}_{\ell,h}$ when the evaluation dataset is fixed. A larger value means that the head assigns more attention from the query to task-relevant context.

LDscore. We introduce LDscore to measure local-streaming tendency: the extent to which a head concentrates its non-sink attention within a recent context window. For a long-context sample x, let $\mathcal{U}(x)$ denote the evaluated query-token positions. We use zero-based token indices and exclude position 0, which often serves as an attention sink. For $u \in \mathcal{U}(x)$, recent-window size R, layer ℓ, and Query head h, the token-level score is

$$\mathrm{LDscore}_{\ell,h}(u; R) = \frac{\sum_{v=\max(1,u-R+1)}^{u} A_{\ell,h}(u, v)}{\sum_{v=1}^{u} A_{\ell,h}(u, v)}. \tag{17}$$

This is the fraction of non-sink attention mass assigned to the recent window. Averaging over the evaluated query positions gives the sample-level score:

$$\mathrm{LDscore}_{\ell,h}(x; R) = \frac{1}{|\mathcal{U}(x)|} \sum_{u\in\mathcal{U}(x)} \mathrm{LDscore}_{\ell,h}(u; R), \tag{18}$$

For an evaluation dataset $\mathcal{X}$, the head-level score is

$$\mathrm{LDscore}_{\ell,h,\mathcal{X}}(R) = \frac{1}{|\mathcal{X}|} \sum_{x\in\mathcal{X}} \mathrm{LDscore}_{\ell,h}(x; R). \tag{19}$$

We abbreviate $\mathrm{LDscore}_{\ell,h,\mathcal{X}}(R)$ as $\mathrm{LDscore}_{\ell,h}$ when the evaluation dataset and window size are fixed. A larger value means that a greater fraction of the head's non-sink attention lies within the recent context window.

Measurement and visualization settings. All behavioral measurements are computed at Query-head granularity. Inputs are truncated to 32,768 tokens for the Qwen3 series and Llama3.1-8B-Instruct, and to 8,192 tokens for Gemma-7B. The 32,768-token setting matches Qwen3's long-context pretraining length (Yang et al., 2025a). For QRscore, we preserve the relative position of the gold evidence span whenever a sample is truncated. For LDscore, we sample 100 long-context sequences from the `book-65536` subset of ProLong (Gao et al., 2025), tokenize them with the corresponding model tokenizer, truncate each sequence to the model-specific length above, and evaluate the final 128 retained query positions. For visualization, each model–metric matrix is min–max normalized independently; the analyses use the original scores.

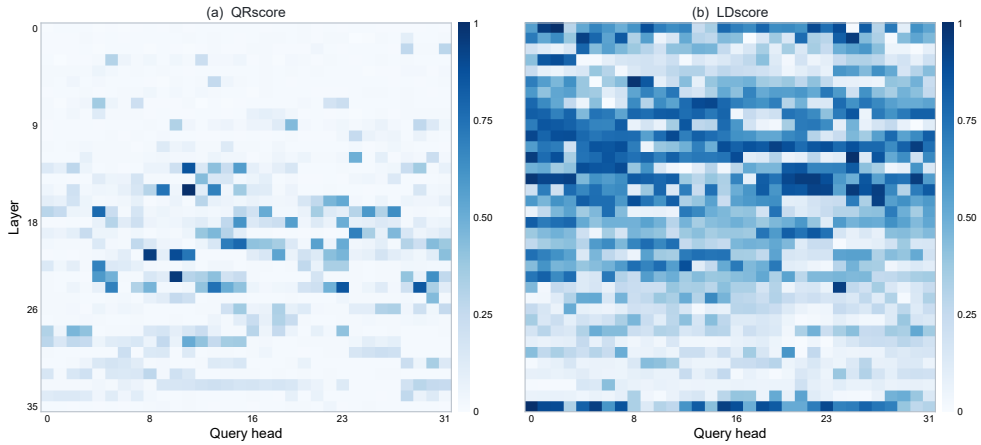

Figure 7: **Qwen3-4B behavioral tendencies.** **(a)** Retrieval tendency measured by QRscore. **(b)** Local-streaming tendency measured by LDscore. Rows denote layers and columns Query heads; each metric is independently normalized for visualization.

A.2 Qwen3-4B Behavioral Patterns

Figure 7 compares retrieval and local-streaming tendencies in Qwen3-4B. Consistent with Figure 1a, retrieval tendency is sparse, negligible at the boundary layers, and concentrated in the middle layers. The first and last layers instead show strong local-streaming tendency, and the two tendencies are approximately complementary across parts of the middle layers. This relation does not hold throughout the network: in several early layers, including layers 1–6, retrieval tendency is almost absent without a correspondingly strong or structured local-streaming pattern; in later layers before the final layer, including layers 25–34, both tendencies are weak and broadly distributed. QRscore and LDscore therefore identify behavioral anchors but do not yield a complete taxonomy.

A.3 Cross-Scale and Cross-Family Behavioral Consistency

Qwen3-1.7B and Qwen3-8B test within-family consistency across scale (Yang et al., 2025a), while Llama3.1-8B-Instruct (Grattafiori et al., 2024) and Gemma-7B (Team et al., 2024) provide cross-family comparisons. Figure 8 shows consistent retrieval and local-streaming trends across the three Qwen3 scales and similar layer-specific organization in both other model families. The within-family agreement may partly reflect the shared architecture and training pipeline. Qwen3 reports strong-to-weak logit distillation for smaller post-trained models, but not explicit attention-head alignment; we therefore leave the source of the close head-level agreement open.

B Additional RFIS and RPD Analysis

B.1 Measurement Settings

For the pretrained-model RFIS and RPD measurements, we sample 100 long-context sequences from the `book-65536` subset of ProLong (Gao et al., 2025) and tokenize them with the corresponding model tokenizer. Sequences are truncated to 32,768 tokens for the Qwen3 series and Llama3.1-8B-Instruct, and to 8,192 tokens for Gemma-7B. The controlled 380M Transformers are evaluated on 2,048-token sequences. In all cases, both metrics are evaluated over the final 128 retained query positions.

We cache the pre-RoPE Q/K vectors, reconstruct the complete logits in equation 7, and perform the RFIS and RPD interventions separately for every layer, Query head, and frequency. The visible-key set is unchanged by either intervention. Under GQA, Query heads that share a KV head are still measured independently because they retain distinct queries, frequency contributions, and complete

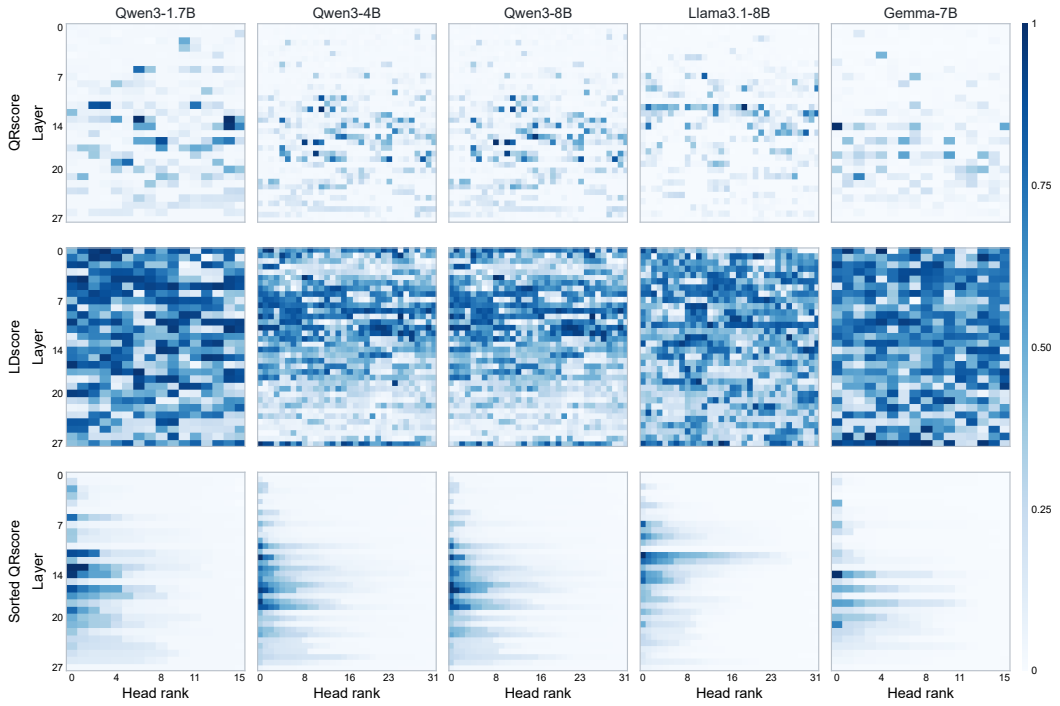

Figure 8: **Behavioral tendencies across model scales and families.** Columns show Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Llama3.1-8B-Instruct, and Gemma-7B. Rows show retrieval tendency (QRscore), local-streaming tendency (LDscore), and retrieval tendency sorted in descending order within each layer. The first two rows preserve the original Query-head order; each model–metric matrix is independently normalized for visualization.

attention distributions. RFIS and RPD use the same samples and query positions, and all analyses use their raw bounded scores without head-wise normalization.

Each RFIS/RPD panel is visualized on its own raw zero-to-maximum scale. To resolve variation near zero, the heatmap color assigned to a normalized raw score $t = s/s_{\max}$ follows the monotone display mapping $t \mapsto t^{0.20}$. The colorbar uses the same mapping with a uniformly distributed palette and raw-score ticks sampled more densely near zero. This transformation affects visualization only and does not modify the reported scores or analyses.

B.2 Layer-Mean Frequency Structure in Qwen3-4B

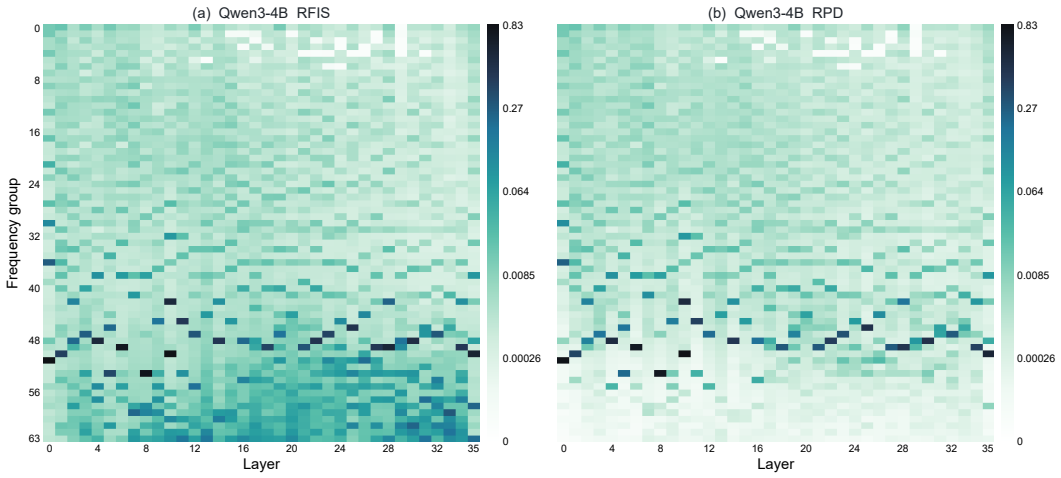

Figure 9: **Layer-mean RFIS and RPD in Qwen3-4B. (a)** RFIS; **(b)** RPD. Scores are averaged over Query heads, with frequencies descending from high to low; panels use independent zero-to-maximum scales. Multiple GPBand-like bands vary with depth, unlike the single layer-invariant band in controlled Transformers.

Figure 9 complements the head-level plots in the main text by exposing the aggregate band locations across all layers. The RFIS and RPD means contain multiple separated bands rather than the single sharp ridge observed under controlled training. A plausible explanation is Qwen3's multi-stage pretraining, whose final stage increases the sequence length from $4{,}096$ to $32{,}768$ (Yang et al., 2025a): different layers may retain positional scales learned at different stages. Because the public training history does not identify a unique effective length for every layer, these dispersed bands are treated as GPBand-like structures rather than used to estimate a single training length. This motivates the controlled experiments below, where the training length and RoPE base are known.

B.3 Cross-Family Mechanism-Level Consistency

Figure 10 provides a mechanism-level cross-family check. In both Llama3.1-8B-Instruct layers, retrieval-associated heads emphasize lower frequencies with weak positional dependence, whereas positional heads show similar RFIS and RPD at the salient mid-low-frequency band and above. The LD strips again do not yield an alternative complete taxonomy. Llama3.1 therefore reproduces the RoPE-frequency dependence underlying the retrieval–positional differentiation found in Qwen3-4B.

Implication for GQA. Figures 10 and 2 reveal more than mixed functions within the same KV group. Because Query heads in a GQA group share the same Key frequency coordinates, their frequency dependencies interfere through this common representation: some individual heads simultaneously exhibit characteristics of both the below-band retrieval pattern and the band-and-above positional pattern. The dominant two functions remain identifiable, but their signatures are often entangled rather than cleanly separated. Future work should quantify the performance effects of this interference and study function-aligned KV grouping or other ways to decouple the two dependencies.

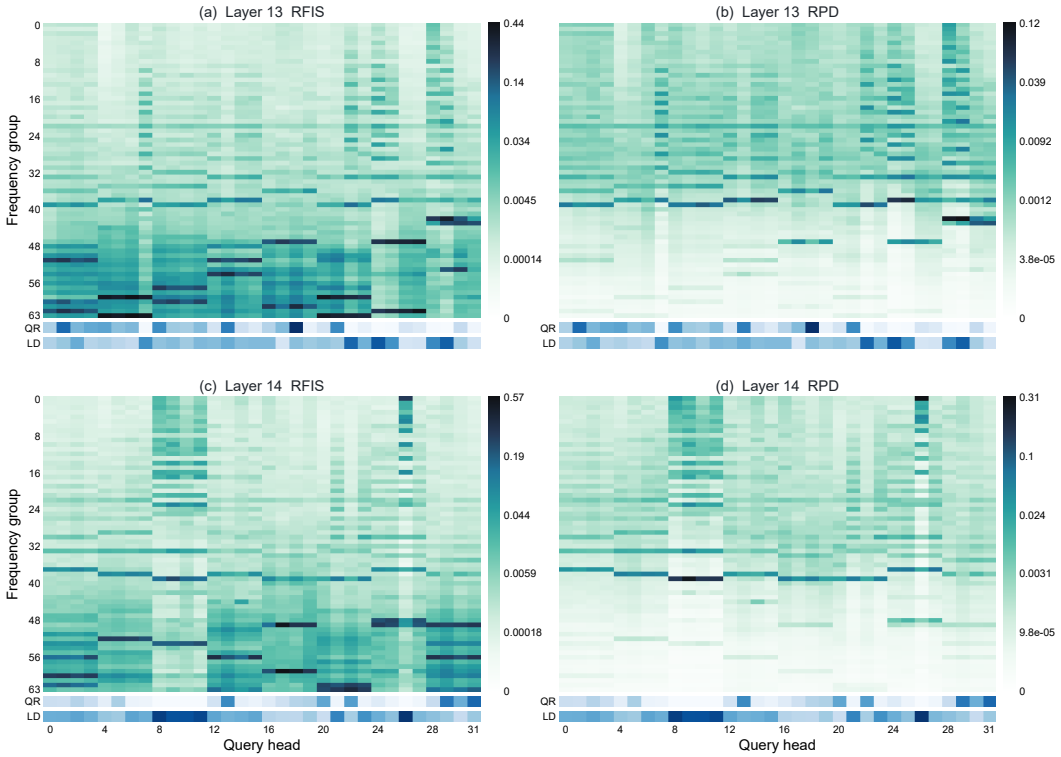

Figure 10: **Cross-family RoPE-frequency dependence in Llama3.1-8B-Instruct.** Rows show layers 13 and 14; columns show RFIS and RPD. Aligned QR and LD strips report retrieval and local-streaming tendencies. Frequencies descend from high to low, and each RFIS/RPD panel uses its own raw zero-to-maximum scale.

B.4 Controlled Analysis of GPBand

Configuration and observed shifts. We compare three matched approximately 380M-parameter Transformers with QKNorm with head dimension $d_h = 64$: two are trained at length 2048 with RoPE bases $\theta = 10^4$ and 10^6, and the third is trained at length 4096 with $\theta = 10^4$. Architecture, tokenizer, data, and optimization are otherwise fixed. Figure 11 shows that changing θ moves the band from groups 21–22 to 14–15, while doubling the training length moves it to approximately group 24.

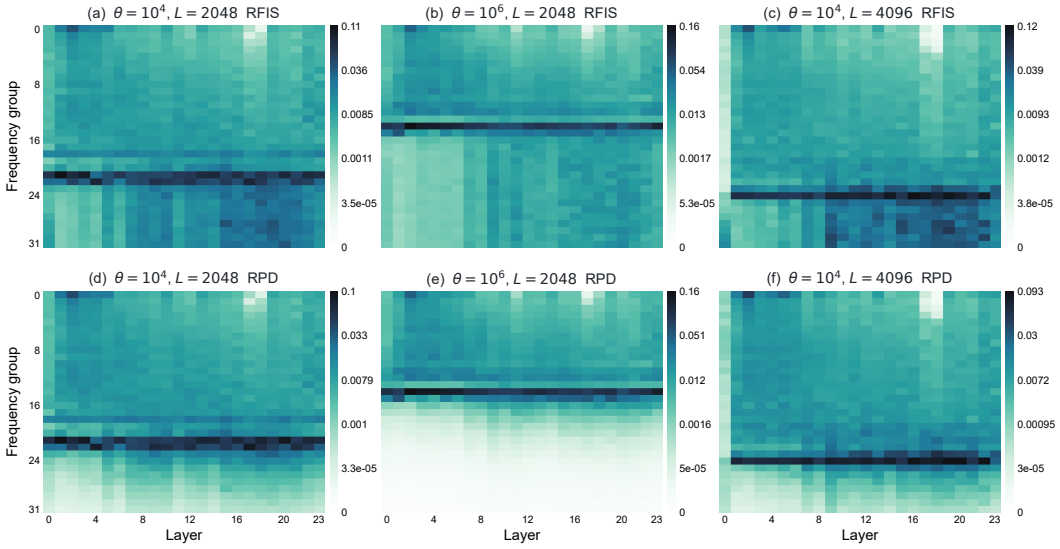

Figure 11: **Controlled RoPE-base and training-length comparisons with QKNorm.** Rows show layer-mean RFIS and RPD. Columns show $(\theta, L_{\text{train}}) = (10^4, 2048)$, $(10^6, 2048)$, and $(10^4, 4096)$. Each panel uses an independent raw zero-to-maximum scale.

With $F = d_h/2 = 32$ two-dimensional frequency groups, group r has angular frequency and rotation period

$$\omega_r = \theta^{-r/F}, \qquad \lambda_r = \frac{2\pi}{\omega_r} = 2\pi\theta^{r/F}. \tag{20}$$

Table 3 converts the observed bands into positional scales. The coordinate index changes with θ and training length, while the corresponding periods remain tied to the training-length scale.

Table 3: **Rotation periods at the controlled GPBands.**

θ	L_{train}	Group r	Period λ_r	$\lambda_r/L_{\text{train}}$
10^4	2048	21	2650	1.29
10^4	2048	22	3533	1.73
10^6	2048	14	2650	1.29
10^6	2048	15	4080	1.99
10^4	4096	24	6283	1.53

Connection to Frequency Bands. Oka et al. (2026) locate a high-norm band from two-dimensional Q/K norms and explain it through the positional variation available to RoPE over the training length. Their appendix jointly analyzes the sine and cosine coordinates through the full covariance

$$\mathbf{\Sigma}(x) = \text{Cov}_{m\sim\text{Unif}[0,L_{\text{train}}]}\left[\begin{pmatrix}\cos(m\omega)\\ \sin(m\omega)\end{pmatrix}\right] = \begin{pmatrix}\Sigma_{11}(x) & \Sigma_{12}(x)\\ \Sigma_{12}(x) & \Sigma_{22}(x)\end{pmatrix}, \quad x = \omega L_{\text{train}}, \tag{21}$$

where

$$\Sigma_{11}(x) = \frac{1}{2} + \frac{\sin(2x)}{4x} - \left(\frac{\sin x}{x}\right)^2,$$
$$\Sigma_{22}(x) = \frac{1}{2} - \frac{\sin(2x)}{4x} - \left(\frac{1-\cos x}{x}\right)^2, \tag{22}$$
$$\Sigma_{12}(x) = \frac{1-\cos(2x)}{4x} - \frac{\sin x}{x}\frac{1-\cos x}{x}.$$

Under a fixed coefficient-norm budget, the maximum centered variation is determined by $\lambda_{\max}(\mathbf{\Sigma}(x))$. Maximizing this quantity gives $x^\star \approx 4.493409$ and hence

$$r^\star = F \log_\theta\left(\frac{L_{\text{train}}}{x^\star}\right), \qquad \lambda^\star = \frac{2\pi L_{\text{train}}}{x^\star} \approx 1.398 L_{\text{train}}. \tag{23}$$

For $L_{\text{train}} = 2048$, this predicts $r^\star = 21.27$ at $\theta = 10^4$ and $r^\star = 14.18$ at $\theta = 10^6$, with period 2864 in both cases, matching groups 21–22 and 14–15. For $L_{\text{train}} = 4096$ and $\theta = 10^4$, it predicts $r^\star = 23.68$ and period 5727, close to observed group 24 with period 6283. RFIS therefore recovers the band through its effect on the complete attention distribution, while RPD establishes that the same band is specifically position dependent. Layer 0 of the length-4096 model appears unusually retrieval-like, which may be a small-model boundary-layer effect; it does not affect the layer-specific distribution consistently observed in Qwen3 and Llama3.1.

Effect of QKNorm. The norm-defined empirical band in Oka et al. (2026) differs from the theoretical index by an approximately 1.1 factor in several pretrained models. Figure 12 shows that removing QKNorm shifts the $\theta = 10^4$ band from groups 21–22 toward group 23 and the $\theta = 10^6$ band from groups 14–15 toward group 15–16, after layer 0.

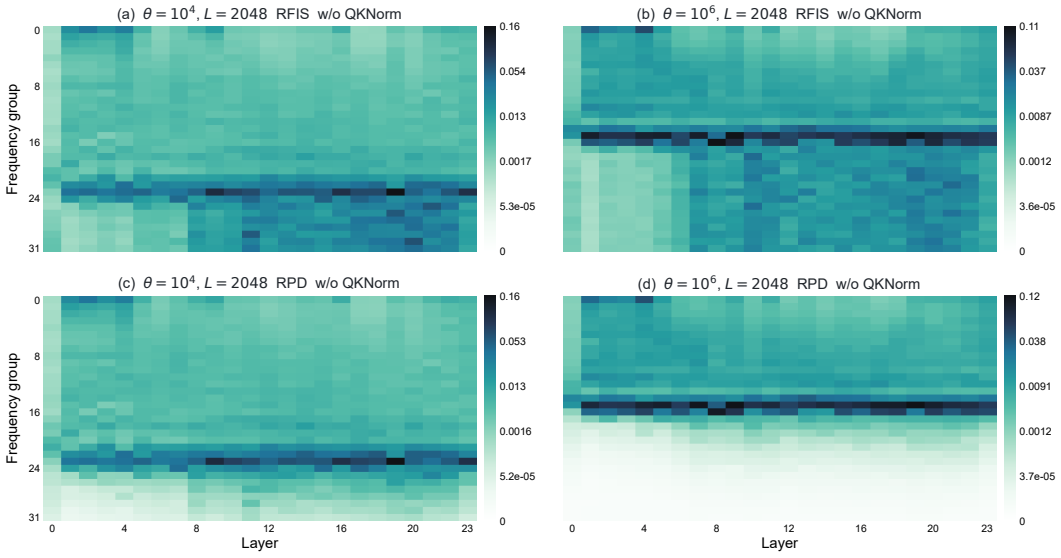

Figure 12: **Controlled models without QKNorm.** Rows show layer-mean RFIS and RPD at training length 2048; columns show $\theta = 10^4$ and 10^6. Relative to Figure 3, the bands shift toward lower frequencies.

This suggests one possible explanation for part of this offset: without QKNorm, the growth of Q/K norms may need to be compensated for by shifting the band toward lower frequencies and smaller rotary angles to maintain stable global positional representations. This interpretation remains a hypothesis; direct norm trajectories would be needed to verify the causal training dynamics.

B.5 Gemma-7B and the Limits of Norm-Based Analysis

Gemma-7B provides a useful comparison with the norm-based analysis of Barbero et al. (2025). The RFIS/RPD detections in Figure 13 show that its dominant position-dependent Band extends

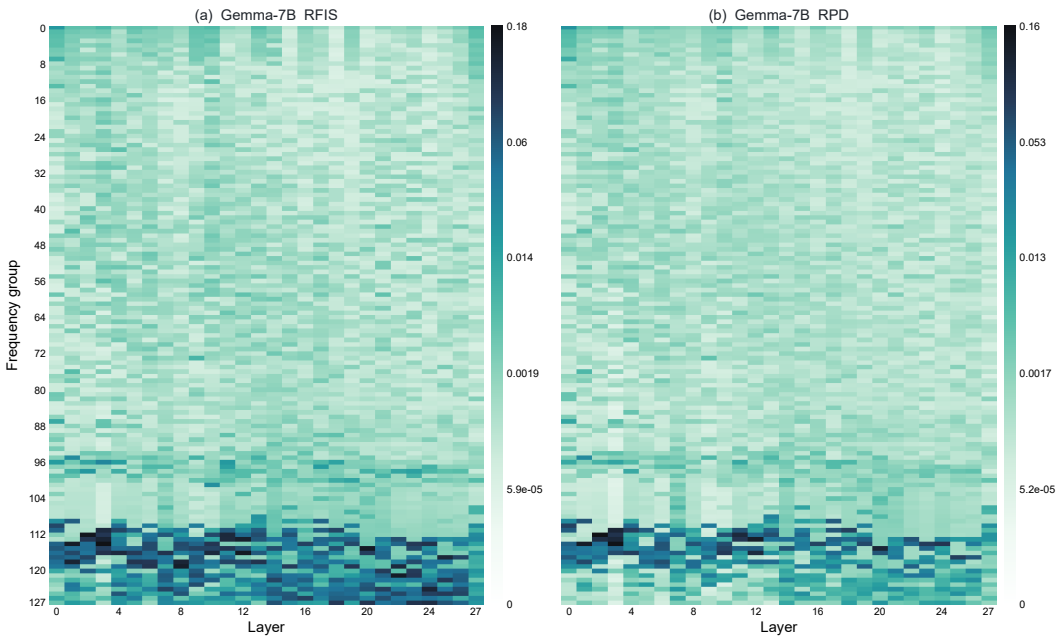

Figure 13: **Layer-mean RFIS and RPD in Gemma-7B.** The dominant position-dependent structure lies near groups 110–120, below the training-length prediction near group 104; a weaker band appears near group 96. Panels use independent raw zero-to-maximum scales.

unusually far into the low-frequency range. With $F = 128$, $\theta = 10^4$, and an approximately 8000-token training length, equation 23 predicts $r^\star \approx 104.0$ and $\lambda^\star \approx 11{,}186$. Groups 100 and 110 have periods 8379 and 17,206, respectively, so the expected training-length-related band lies between them. RPD, however, remains strong through approximately group 120, whose period is 35,333. This extension may reflect the absence of QKNorm or an undocumented training pipeline. A weaker band near group 96, with period $6283 < 8000$, is also visible, but its role remains unclear.

This result changes the interpretation of the low-frequency dependence reported by Barbero et al. (2025): a substantial portion of the region identified through Q/K norms is highly position dependent rather than position-invariant retrieval structure. Frequency location and norm magnitude therefore cannot directly determine whether a RoPE channel depends on positional modulation; RPD is required for that distinction.

Figure 14 nevertheless shows the same two dominant dependence patterns in Gemma-7B layers 14 and 16: retrieval-associated heads emphasize lower frequencies with comparatively weak RPD, whereas positional heads show matched RFIS/RPD activity near the shifted GPBand and above. Because the boundary lies close to the lowest-frequency end and the head contains 128 frequency groups, each individual contribution is visually diluted, making the positional pattern less distinct than in Qwen3 or Llama3.1. The shifted and extended boundary therefore remains unresolved, but it neither provides a stable third frequency-dependence pattern nor contradicts our central retrieval–positional differentiation result.

C Theoretical Details of RFIS and RPD

Relation to norm-based analysis. Barbero et al. (2025) approximate RoPE-frequency usage using mean per-frequency Q/K norms, motivated by the fact that the Cauchy–Schwarz inequality bounds the magnitude of each frequency's contribution in terms of these norms. Because norms discard directional alignment between Q and K, this magnitude-only approximation introduces error when estimating each frequency's realized contribution to attention. RFIS instead removes the exact contribution of one frequency from the Q/K attention computation while holding all others fixed and measures the resulting change in the complete attention distribution, providing a more faith-

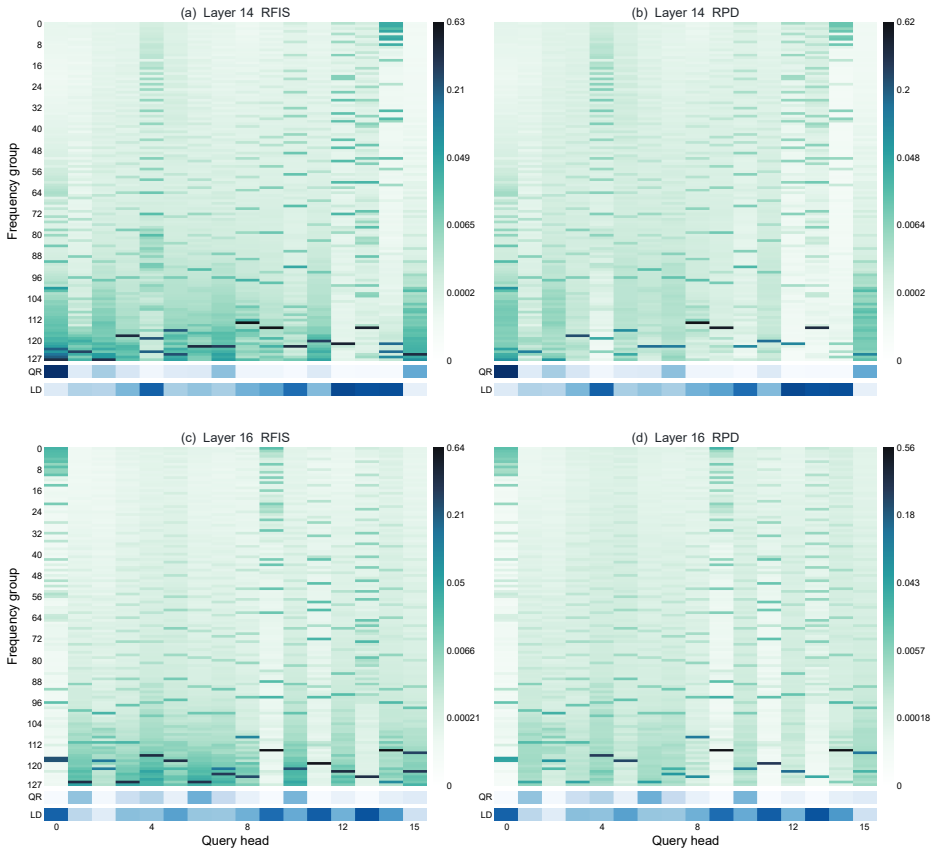

Figure 14: **Head-level RoPE-frequency dependence in Gemma-7B.** Rows show layers 14 and 16; columns show RFIS and RPD with aligned QR and LD strips. Frequencies descend from high to low, and each RFIS/RPD panel uses its own raw zero-to-maximum scale.

ful frequency-importance metric. RPD then applies a further controlled intervention: it leaves the frequency's Q/K vectors unchanged and replaces only its relative RoPE rotation with the identity, thereby isolating dependence on rotary positional modulation, which norm-based analysis cannot measure.

C.1 Boundedness of the normalized JS readout

Throughout the RFIS and RPD definitions, $\log$ denotes the natural logarithm. Let $\mathbf{m} = (\mathbf{p} + \mathbf{q})/2$. Non-negativity follows from the two KL terms in equation 9. For the upper bound, $m_j \geq p_j/2$, so for every $p_j > 0$,

$$\log \frac{p_j}{m_j} \leq \ln 2. \tag{24}$$

Therefore,

$$D_{\mathrm{KL}}(\mathbf{p} \| \mathbf{m}) = \sum_j p_j \log \frac{p_j}{m_j} \leq \ln 2. \tag{25}$$

The same argument applies to $\mathbf{q}$, giving

$$0 \leq D_{\mathrm{JS}}(\mathbf{p}, \mathbf{q}) \leq \ln 2. \tag{26}$$

Equality at the upper bound requires disjoint supports. A softmax with finite valid logits assigns positive probability throughout the common support, so attention distributions generally approach rather than attain the normalized value 1. The value $\ln 2$ therefore follows only from the chosen logarithm base. With base-b logarithms, the upper bound and normalization constant would both be $\log_b 2$, leaving the normalized RFIS and RPD scores unchanged.

D Additional Experimental Details

Pretraining. We train matched models from scratch on the `sample-100BT` subset of FineWeb-Edu (Penedo et al., 2024) with the Mistral tokenizer (vocabulary size: 32,000) and a sequence length of 2048. The approximately 380M and 1.4B models are trained for 15B and 100B tokens with global batch sizes of 256 and 1,024, respectively, using standard mixed-precision training. Within each scale, all architecture comparisons use the same data order, optimizer schedule, and evaluation protocol. Attention variants are width matched by holding the total Q/K/V dimension per layer at 1024 and 2048, respectively. At 380M, NoPE-FA heads use Q/K/V dimension 64 and GDN heads use 128; at 1.4B, both use 128. Ratio labels in Section 4 refer to the share of total head dimension. The small parameter-count differences arise from GDN's additional decay and gating parameters; because the capacity-defining Q/K/V widths are matched, these auxiliary parameters do not materially affect the fairness of the comparison. All models use fused AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-15}$, weight decay 0.1, and a peak learning rate of 3×10^{-4}. We apply 1024 warmup steps followed by cosine decay to 10% of the peak rate and clip the gradient norm at 1.0.

Baselines. Transformer uses RoPE FA throughout, GDN uses only LA, and Inter alternates three GDN layers with one NoPE FA layer. These baselines isolate homogeneous FA, homogeneous LA, and coarse layer-wise hybridization from HwH's head-wise, layer-specific allocation.

Evaluation. All evaluations use LM Evaluation Harness (Sutawika et al., 2023). Perplexity uses WikiText (Merity et al., 2017) and the OpenAI-formatted LAMBADA benchmark (Paperno et al., 2016). Commonsense reasoning uses LAMBADA, HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), ARC-Easy/Challenge (Clark et al., 2018), WinoGrande (Sakaguchi et al., 2019), and OpenBookQA (Mihaylov et al., 2018). Real-world retrieval uses FDA and SWDE (Arora et al., 2024a), SQuAD (Rajpurkar et al., 2018), Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and DROP (Dua et al., 2019), using the cloze-completion variants of Arora et al. (2024a;b). Synthetic retrieval uses RULER NIAH-Single-1/2/3 (Hsieh et al., 2024). The dense NIAH-Single-2 evaluation follows the length–depth-like protocol of gkamradt (2026) with the fixed-depth RULER construction, the Paul Graham Essays haystack, and 128 generation tokens. The grid contains context-length limits $1024, 1536, \ldots, 8192$ and depth percentages $0\%, 10\%, \ldots, 100\%$, where 0% is the shallowest insertion at the sequence end and 100% the deepest at the sequence beginning. We sample 10 examples for each of the 15×11 cells, totaling 1650

examples per model. Heatmaps place 100% depth at the top and 0% at the bottom; overlaid curves average accuracy across the 11 depths at each length and use the same 0–100 vertical scale.

Short-length instability at small scale. The 380M Transformer and HwH-std-r exhibit anomalously poor NIAH-Single-2 performance at 1K. Rather than returning the seven-digit target, their specific erroneous responses include `_saltimbocca_`, `_saltintesta_`, and the underscore-only blank `_____`. These are not simply repeated phrases that dominate the haystack: for example, `_saltimbocca_` occurs only once in the corresponding sample, roughly 10–20% into the sequence, yet is selected as the answer. This behavior suggests that the affected small-scale architectures can lock onto a salient distractor or fall into a blank-response mode even within short contexts, reflecting limited short-sequence robustness rather than length-extrapolation behavior. The 1.4B Transformer remains stable at both 1K and 2K, further showing that the phenomenon is scale- and configuration-dependent.

Controlled GPBand models. The controlled 380M RoPE Transformers follow the same data, tokenizer, model scale, and optimization hyperparameters as the 380M pretraining runs above. They vary only the factors under study: the RoPE base, training sequence length, and, for the designated variants, the use of QKNorm.

E Additional HwH Configuration Details

Tokenizer. All models use the same Mistral tokenizer with a vocabulary size of 32,000 across model scales and architecture variants.

Table 4 records the exact distinctions among the 380M ablation models. "Outer layers" denotes the shallow and deep layers outside the middle half, excluding the first and last layers. HwH-swa uses sink-free RoPE SWA with window 128 as its local component. All other variants use GDN.

Table 4: Configurations of the HwH ablation models. Ratios denote global-component:local-component Q/K/V dimensions.

Model	Global	Local	First/last	Middle half	Outer layers
HwH-uni	NoPE FA	GDN	1:3	1:3	1:3
HwH-rmfl	NoPE FA	GDN	all local	1:3	1:3
HwH-std	NoPE FA	GDN	all local	1:3	1:7
HwH-swa	NoPE FA	SWA	all local	1:3	1:7
HwH-std-r	RoPE FA	GDN	all local	1:3	1:7

E.1 Detailed Ablation Results

Tables 5–8 provide the task-level ablation results, and Figure 15 consolidates the complete length–depth grids for all 380M baselines and ablations.

Table 5: Language modeling for the functional-component ablation at 380M. Metrics and formatting follow Table 1;

Model	Wiki ↓	LAMB. ↓	LAMB.	HellaS.	PIQA	ARC-E	ARC-C	WinoG.	OBQA	Avg.
HwH-std	27.41	**34.84**	**32.76**	39.93	**67.57**	**57.87**	28.16	50.75	**22.40**	**42.78**
HwH-swa	28.97	38.54	**32.76**	39.11	66.81	56.19	27.22	51.85	21.00	42.13
HwH-std-r	**27.32**	38.15	31.01	**40.01**	66.87	57.62	**29.52**	**52.64**	20.80	42.64



Figure 15: **Complete** 380**M NIAH-Single-2 results.** Rows show baselines, component ablations, and allocation ablations, with HwH-std repeated as reference. Depth runs from 100% (deepest; top) to 0% (shallowest; bottom); white curves are depth averages on the same scale.

Table 6: Real-world retrieval for the functional-component ablation at 380M.

Model	FDA	SWDE	SQuAD	NQ	TriviaQA	DROP	Avg.
HwH-std	24.14	**31.05**	**32.74**	17.29	**49.17**	19.45	28.98
HwH-swa	**30.22**	26.73	31.90	**17.58**	47.04	**22.47**	**29.32**
HwH-std-r	6.53	21.15	31.74	14.06	48.93	19.60	23.67

Table 7: Language modeling for the layer-specific allocation ablation at 380M. Metrics and formatting follow Table 1;

Model	Wiki ↓	LAMB. ↓	LAMB.	HellaS.	PIQA	ARC-E	ARC-C	WinoG.	OBQA	Avg.
HwH-uni	27.63	34.38	32.66	**40.18**	66.97	**59.01**	27.39	52.88	21.40	**42.93**
HwH-rmfl	27.94	**33.85**	**33.55**	39.63	66.76	57.79	27.65	**53.35**	21.40	42.88
HwH-std	**27.41**	34.84	32.76	39.93	**67.57**	57.87	**28.16**	50.75	**22.40**	42.78

Table 8: Real-world retrieval for the layer-specific allocation ablation at 380M.

Model	FDA	SWDE	SQuAD	NQ	TriviaQA	DROP	Avg.
HwH-uni	21.78	27.18	31.67	15.36	48.05	18.54	27.10
HwH-rmfl	19.96	24.75	**33.34**	16.69	48.58	18.59	26.99
HwH-std	**24.14**	**31.05**	32.74	**17.29**	**49.17**	**19.45**	**28.98**