THE FATES' SPINDLE: ROTATION REPRESENTATION EMERGES FOR RELATIVE POSITION IN TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models rely on relative positional information across diverse tasks, while the mechanisms underlying relative position reasoning have yet to be elucidated. Using a controlled synthetic task that isolates relative position retrieval, we systematically study Transformers across positional encoding schemes and uncover a unified geometric mechanism. Across NoPE, ALiBi, T5 relative bias, KERPLE, and learned absolute embeddings, models converge to nearly identical geometry: keys and queries collapse into an approximately two dimensional subspace and trace a uniform rotation completing one cycle over the training context. This structure is equivalent to generalized RoPE, with all active channels sharing a frequency of approximately $2\pi/L$ per position, where L is the training context length. We further show that this single frequency representation is theoretical inferior to suitable multifrequency alternatives, explaining the advantage of richer frequency components in RoPE and related representations. Guided by this analysis, we evaluate a new frequency allocation scheme, providing support for the framework and its practical utility. Our findings provide a understanding of how Transformers learn relative position representations, establish synthetic tasks and geometric analysis as a principled approach to investigating positional capabilities, and inform future positional encoding design.

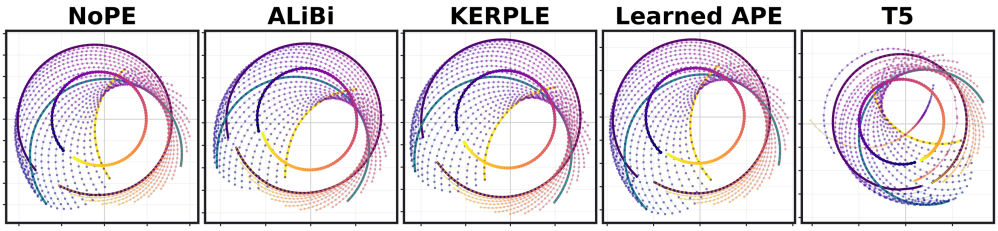

Figure 1: **The Fates' Spindle.** Joint PCA of centered, position averaged keys and marker queries. Colors indicate target aligned position, inner ring consists of keys, while the three bold lines in the outer circle indicate the three markers (m=1, 10, 20).

1 INTRODUCTION

Large language models have achieved remarkable success across a broad range of natural language tasks, where the ability to reason over relative positional relationships plays a fundamental role. Many dependencies in language are inherently relational rather than absolute: a model may need to retrieve a token appearing several positions earlier, associate an entity with its previous mention, or propagate information across a specific displacement. Such operations require the model to represent and compare relative distances between tokens, making relative position retrieval a core computational primitive of Transformers (Vaswani et al., 2017; Shaw et al., 2018).

Positional encodings supply substantially different positional signals to Transformers: learned absolute positional embeddings (APE) (Devlin et al., 2019), relative position representations (Shaw et al., 2018), rotary positional embeddings (RoPE) (Su et al., 2024), and relative attention biases, including T5 (Raffel et al., 2020), ALiBi (Press et al., 2021), and KERPLE (Chi et al., 2022). Causal

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.

Transformers without explicit positional embeddings (NoPE) can also acquire positional information (Haviv et al., 2022; Kazemnejad et al., 2023). However, whether these different encodings induce distinct internal mechanisms for relative position retrieval, or converge toward a shared computational solution, remains unclear.

Directly studying this question in natural language is challenging, as semantic structure, long range dependencies, and diverse linguistic phenomena introduce substantial confounding factors. To isolate the essential mechanism of relative position retrieval, we introduce a synthetic task, XoM, that abstracts relative position retrieval into its simplest form. Each sequence contains a marker associated with a target offset, and the model must retrieve the token located exactly that offset away from the marker. Since the marker can appear at arbitrary positions, successful prediction requires learning a translation invariant relative rule rather than memorizing absolute positions.

We first analyze how Transformers equipped with RoPE solve XoM. RoPE explicitly rotates query and key representations according to absolute positions (Su et al., 2024). Therefore, the attention score depends on their phase difference, allowing the model to retrieve the token at a desired relative offset by aligning the query phase with the corresponding key phase, consistent with prior analyses of position-selective RoPE attention (Barbero et al., 2024). This also provides a direct geometric interpretation of how relative position matching can be implemented in other models.

Surprisingly, across NoPE, ALiBi, T5-style relative bias, KERPLE, and learned APE, the internal representations at the critical retrieval layer converge to nearly the same geometric structure. Position averaged keys become approximately two dimensional and trace near circular trajectories dominated by a single frequency corresponding to one rotation over the training context length. Moreover, the marker query follows an almost identical trajectory, enabling relative matching through "phase" differences. This reveals that Transformers can spontaneously construct a generalized RoPE-like representation even when no rotary positional mechanism is provided. The emergent representation is a solution discovered by optimization to satisfy the geometric requirements of relative position retrieval. And the convergence evokes the Fates' spindle: optimization appears to weave distinct positional mechanisms into a common rotational geometry under the demands of relative retrieval.

This observation raises several fundamental questions:

1. *Why does relative position matching naturally induce a circular representation?*
2. *Why do models converge to a low dimensional single frequency solution despite the availability of higher dimensional representations?*
3. *Given that richer multi frequency representations can provide stronger positional discrimination, what is the optimal frequency allocation for rotary position representations?*

We answer these questions from a geometric perspective. First, we formulate relative position retrieval as a geometric separation problem and show that the two dimensional circular representation maximizes positional discrimination under the minimum transition constraint. Second, we analyze the optimization dynamics and identify a measurable concentration mechanism in the Q/K projection weights that promotes energy accumulation into a small number of positional directions, explaining the emergence of low rank representations. Finally, we revisit RoPE from the perspective of optimal frequency allocation. Although RoPE provides multi frequency representations and substantially improves over the learned single frequency solution, its manually designed frequency schedule is not optimal for relative position retrieval. We introduce a new frequency allocation rule whose induced representation asymptotically approaches the task's optimal solution under a sufficient asymptotic condition.

To empirically validate the preceding geometric analysis and the proposed analytical framework, we evaluated the geometrically optimized RoPE frequency allocation scheme on the synthetic task and in realistic language modeling settings. The results corroborate our analysis and demonstrate that a framework grounded in synthetic tasks and geometry offers a viable path toward understanding and improving the positional capabilities of large language models.

Our contributions are summarized as follows:

- **A unified mechanism for relative position learning.** Through experiments on five representative non-RoPE positional encodings, we reveal that Transformers independently converge to a gener-

alized RoPE-like representation: a low dimensional single frequency circular trajectory shared by keys and queries at the retrieval layer.

- **A geometric Interpretation of relative position representations.** We establish a connection between relative position retrieval and geometric separation, characterize the optimal solutions under different dimensional constraints.
- **A principled RoPE frequency allocation strategy.** Guided by the derived geometric principles, we develop a new frequency allocation method for RoPE to empirically validate our analytical framework. The results support the viability of combining synthetic tasks with geometric analysis and provide meaningful guidance for improving the positional capabilities of language models.

2 PRELIMINARIES

2.1 XOM: A SYNTHETIC RELATIVE POSITION RETRIEVAL TASK

An input sequence $x = (x_0, \ldots, x_{L-1})$ contains exactly one special marker m at position i. Each marker type specifies a fixed offset $d_m > 0$, and the prediction target at the marker position is

$$y_i = x_t, \qquad t = i - d_m.$$

The cross entropy loss is computed only between the output at marker positions and y_i. Multiple markers are used as real language tasks contain tokens with diverse relative positional relationships. A marker placement is valid only when $i \geq d_m$, and a marker with offset three in

$$[\ldots, x_a, x_b, x_c, \boxed{m}, \ldots] \quad \text{requires predicting } x_a.$$

All other tokens in the sequence are sampled uniformly at random. Thus, marker identity specifies where to retrieve, rather than which token to predict. Our main experiments use $L = 100$ and a vocabulary of twenty marker types, with

$$d_m = 2 + 4(m-1), \qquad m \in \{1, \ldots, 20\}.$$

The model has 3 Transformer layers, 2 attention heads per layer, model width 64, and head width 32. Each configuration is trained for 210 epochs using three random seeds. Experiment details and results for additional models exhibiting similar mechanisms are provided in the Appendix.F.

2.2 HOW ROPE ACCOMPLISHES XOM

We specialize RoPE's relative rotation formulation (Su et al., 2024) and the shifted attention construction of Barbero et al. (2024) to relative position retrieval. Let $R(\theta)$ be a plane rotation and $\mathsf{R}(j) = \bigoplus_r R(j\omega_r)$ its block rotation at position j. For pre rotation vectors $\widetilde{q}_m, \widetilde{k}$, its numerator is

$$(\mathsf{R}(i)\widetilde{q}_a)^\top \mathsf{R}(j)\widetilde{k} = \widetilde{q}_a^\top \mathsf{R}(j-i)\widetilde{k}.$$

Marker aligned to target phase. Consider a position independent unit vector u and the marker specific pre rotation query

$$\widetilde{q}_m = \mathsf{R}(-d_m)u.$$

Using u as the pre rotation key gives

$$k_j = \mathsf{R}(j)u, \qquad q_a = \mathsf{R}(i - d_m)u.$$

The query thus matches the phase of the target address $t_a = i - d_m$. The construction applies at every legal marker position.

Token copying via phase alignment. If $w_r = \|u_r\|^2$ is the seed energy in block r, the score becomes

$$q_a^\top k_j = \sum_r w_r \cos[\omega_r(j - t_a)].$$

At the target, all blocks agree and the score is one, and the separation between the target's score and the distractor's score is given by

$$q_a^\top k_{t_a} - q_a^\top k_j = \sum_r w_r \left(1 - \cos[\omega_r(j - t_a)]\right).$$

To be able to obtain as much information of the correct target value as possible in the subsequent weighted value sum, we want the difference between the target's score and the maximum distractor's score to be as large as possible, thus our objective can be written as

$$\max_{u} \min_{j \in J_a \setminus \{t_a\}} q_a^\top (k_{t_a} - k_j).$$

Once this objective value becomes sufficiently large, the weighted value sum then retrieves the target's value representation, from which a suitable readout predicts its token. As we later demonstrate, RoPE indeed operates via this mechanism in practice. Moreover, the above objective applies not only to RoPE, but is also an idealized proxy objective for XoM. The solutions learned by the other five models can be explained through this objective, and the mechanism is also observed in the other five models.

2.3 A GEOMETRIC OBJECTIVE FOR XOM

Let $a = (m, i)$ index a valid retrieval condition, let $t_a = i - d_m$ be its target address, and let $J_a \subseteq \{0, \ldots, N-1\},\ N \leq L$ be its candidate set, including t_a. We maximize the worst query key margin

$$\begin{aligned} \gamma(Q, K) &= \min_{a} \min_{j \in J_a \setminus \{t_a\}} q_a^\top (k_{t_a} - k_j), \\ \text{s.t. } \|q_a\|_2 \leq 1, &\quad \|k_j\|_2 \leq 1 \quad \text{for all } a, j,\ \ q_a, k_j \in R^D. \end{aligned} \tag{1}$$

Here, q_a represents the marker's query at the layer of interest, and k_j represents the key associated with absolute address j. Since retrieval is determined by marker identity and relative position, we analyze an idealized addressing model in which keys depend only on position. This abstracts the mechanism without implying that learned keys are inherently independent of token content. This objective essentially requires sharp discrimination of all relative positions, and is an ideal proxy objective whose connection with the true objective can be established through simplified models.

3 THE LEARNED MECHANISM FOR XOM

3.1 ROPE ACCOMPLISHES XOM VIA PHASE

As shown in the Fig.2, RoPE adjusts the phase difference between the marker's query and the key so that the peak is exactly at the offset position, i.e., this position corresponds exactly to $\cos(0)$ in

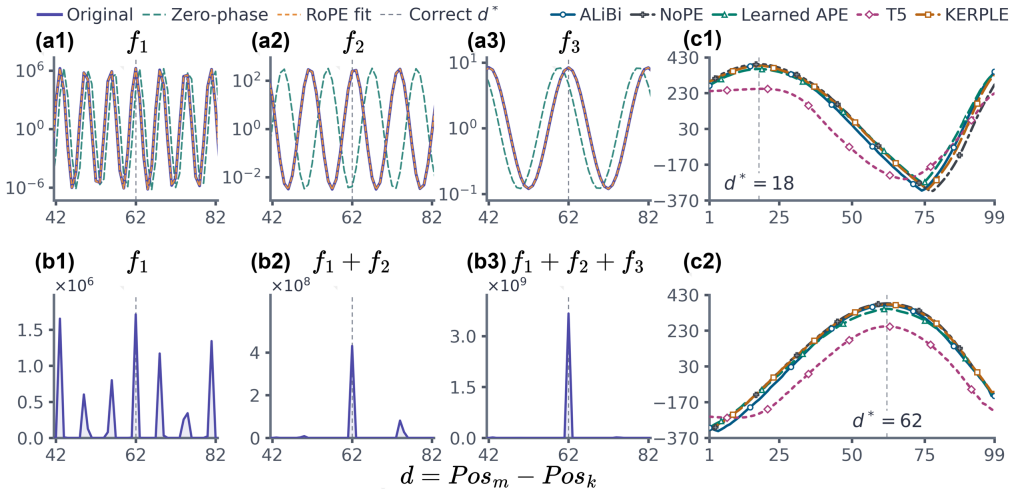

Figure 2: **Phase alignment selects the target offset.** All x axes show query key distance $d = Pos_m - Pos_k$. (a1–a3) Softmax numerators for individual RoPE frequency components, with fitted and zero phase references, y axes are logarithmic. (b1–b3) Softmax numerators combining one, two, and three components, respectively, on linear scales. (c1,c2) Query key scores for five encodings at target offsets 18 and 62, respectively.

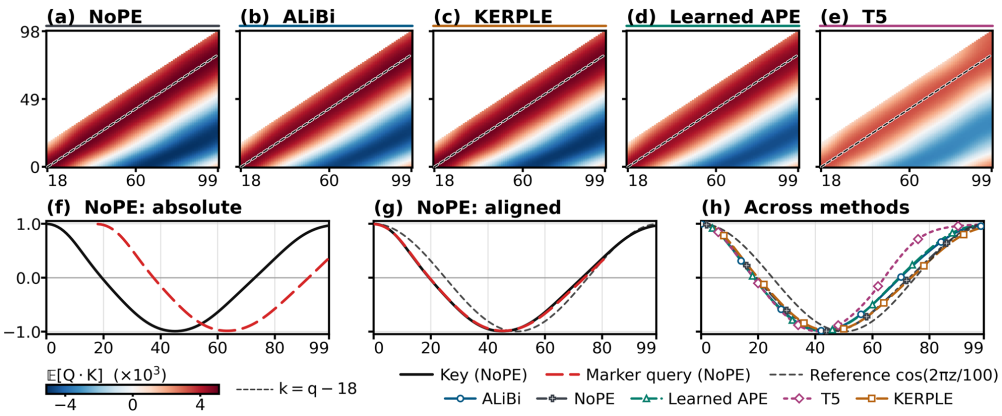

Figure 3: **Encodings align query and key.** (a–e) Mean query key dot products (32 samples) for marker 5 (offset 18), with absolute query position on the x axis and key position on the y axis; dashed lines mark ($k = q - 18$). (f,g) NoPE mean key and query cosine similarities to a reference key at position 0, plotted against absolute position (f) and target aligned position (g: queries shifted by -18). (h) Key cosine similarity versus absolute position across encodings.

the numerator. Thus, we have

$$q_a^\top k_{t_a} = \sum_r w_r \cos[\omega_r \times 0] \geq \sum_r w_r \cos[\omega_r \times (j-t)] = q_a^\top k_j.$$

This provides direct verification of the RoPE mechanism described earlier. We further perform ablations that bypass the attention module or replace its values, confirming that this head is the attention head responsible (Appendix.F). We also show the marker's query and key follow nearly identical angular trajectories (Appendix.F).

3.2 Other encodings also rely on rotation

As illustrated in the Fig.3, all five non-RoPE positional encoding schemes establish the attention peak at the position corresponding to the offset shift within the attention head responsible. The mechanism by which they achieve this translation invariant attention closely mirrors that of RoPE, through the aligned rotation of key and query. As shown in Fig.2, all five encoding schemes have acquired a near cosine pattern highly similar to that of RoPE. All operate by modulating the query "phase" associated with each marker, shifting the peak of the approximately cosine attention curve to the corresponding relative offset. Moreover, as shown in Fig.3, their cosine similarity closely follows a single period cosine over the sequence, further confirms uniform and periodic key rotation. Moreover, the angular variation of the marker's query nearly overlaps with that of the key. Demonstrate that the query reaches the translation invariant attention peak via the same rotation as the key. The above phenomenon is widely present under a variety of different experimental settings (Appendix.F).

Above results reveal two geometric features: First, **the change of key at each step may be consistent and uniform**. Since the norm of the key does not change much (Appendix.F), suggests that the per step transformation of the key is approximately a fixed orthogonal transformation. In particular, the period of this orthogonal transformation is exactly one sequence length. Second, **the per step change of the query at valid positions is approximately consistent with the key**. Consequently, **the transformations of query and key at this critical layer are essentially a generalized RoPE**.

3.3 Other encodings are approximately a single frequency RoPE

We have shown that all five non-RoPE models subject both the key and the marker's query to the same approximately orthogonal transformation (rotation), a finding that amounts to a generalized RoPE. We now show that they have uniformly converged to a two dimensional single frequency RoPE.

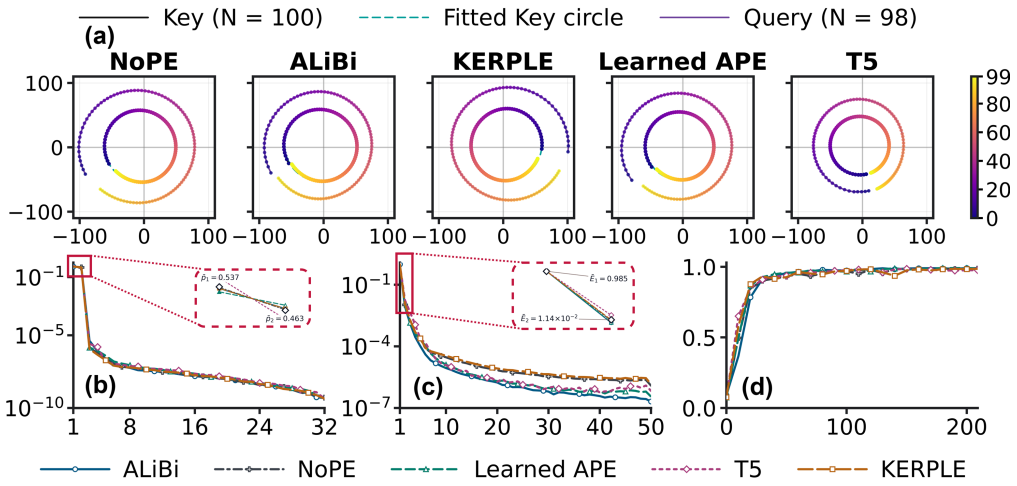

Figure 4: **Encodings learn a two dimensional, single frequency RoPE.** (a) Joint PCA of centered, position averaged keys and marker ($m = 1$ with offset 2) queries. Colors indicate target aligned position, dashed curves are fitted key circles. (b) Key energy fraction $p_j = \sigma_j^2 / \sum_\ell \sigma_\ell^2$ versus singular component index j. (c) Key Fourier energy fraction e_r versus frequency index r, corresponding to angular frequency $2\pi r/100$. (d) Frequency energy e_1 versus training epoch.

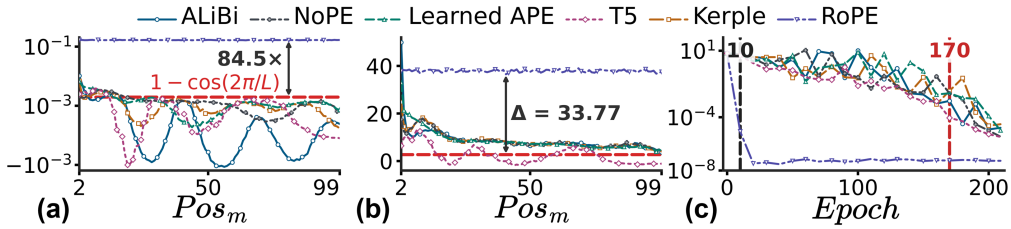

Figure 5: **More frequencies improve positional separation and convergence.** (a) Normalized query key margin against the strongest distractor versus marker position, averaged over 64 samples and excluding additive attention biases. The red dashed line marks the two dimensional benchmark $(1 - \cos(2\pi/100))$. (b) The unnormalized version of (a), the red dashed line denotes the mean of the minimum values across 5 models. (c) Worst marker's test loss versus training epoch.

As illustrated in the Fig.1 and Fig.4, at each position, we average and center the keys to obtain a positional key vector and repeat this procedure for each marker's queries. Joint principal component analysis of the resulting key and query vectors reveals a clear geometric structure at the critical layer: the positional key trajectories closely follow the fitted circles, while the marker's query trajectories trace approximate circles and track changes in key direction.

Spectral analysis corroborates this geometric pattern (Fig. 4). The positional key vectors exhibit two dominant, nearly equal singular values that far exceed all others, indicating an effectively rank two representation. Their energy is overwhelmingly concentrated at frequency $\frac{2\pi}{L}$, exceeding that at every other frequency by more than two orders of magnitude and indicating a single frequency positional feature. And this energy concentration is completed rapidly in early training. Marker queries exhibit analogous singular value and frequency spectra (Appendix.F). Together, these results provide strong evidence that **all five non-RoPE models converge to a two dimensional single frequency generalized RoPE.**

3.4 The Limitation of Single Frequency RoPE

For the XoM objective defined in Section 2.3, the two dimensional single frequency RoPE learned by all five non-RoPE models approximates the optimal solution of minimum transition energy in

two dimension. With unit normalized queries and keys, the optimal objective in two dimension is

$$\max \gamma = 1 - \cos\left(\frac{2\pi}{N}\right). \; E_{\text{path}} = \sum_{j=0}^{N-2} \|k_{j+1} - k_j\|^2 \tag{2}$$

This optimum is attained when the keys are equally spaced on the unit circle and each marker query aligns with its correct offset key. Among these optimal configurations, minimizing the transition energy above yields a constant angular increment of magnitude $2\pi/N$ per position, matching the solution learned by the five non-RoPE models with $N = L$. Consistently, the cosine fit $\cos(2\pi z/L)$ closely tracks the observed positional variation of the keys in Figure 3.

Setting N to the actual sequence length L, we compare the empirical objective values of the five non-RoPE models with that of RoPE in Fig.5. All five models fall below the two dimensional optimum. We compute all models' value using normalized queries and keys for a consistent comparison, negative values in some models arise primarily from their reliance on subsequent bias terms for additional discrimination. RoPE substantially exceeds the two dimensional optimum, enabling sharper positional discrimination at comparable query and key norms and highlighting the benefit of multiple frequencies. Fig.5 also show the unnormalized objective values, and RoPE is still significantly higher than the other five models.

More generally, for $d > 3$, a multifrequency solution exists whose objective value strictly exceeds the two dimensional optimum. Well designed multifrequency solutions such as RoPE achieve substantially larger margins and empirically converge faster, as shown in Figure 5.

Why all five models converge to a low rank single frequency solution rather than a multifrequency alternative remains an important and challenging question. We offer only a brief mathematical explanation. Under the conditions of Theorem C.2, eigenvalue gaps in the optimization matrix can cause Euclidean gradient descent to concentrate energy in eigendirections associated with larger eigenvalues. However, collapse onto the one dimensional leading eigenspace would preclude positional discrimination at the sequence lengths considered, requiring the model to retain at least rank two. Together, these observations suggest that training induced spectral gaps favor concentration in the leading two eigendirections, and the optimal solution in two dimensions takes a single frequency form, yielding a two dimensional single frequency generalized RoPE (Appendix.C).

4 Design a better frequency for RoPE in XoM

4.1 RoPE is not optimal in XoM

Although standard RoPE achieves a larger γ than a single frequency encoding, its frequency allocation remains suboptimal in under the theoretical setting. For odd N and $d = N - 1$, the following attains the optimal $\gamma*$:

$$k_j = \bigoplus_{r=1}^{(N-1)/2} \sqrt{\frac{2}{N-1}} \begin{pmatrix} \cos(j\omega_r) \\ \sin(j\omega_r) \end{pmatrix}, \;\; \omega_r = 2\pi r/N, \;\; \gamma* = \frac{N}{N-1}. \tag{3}$$

By comparison, for sufficiently large N, the standard RoPE frequencies $\omega_r = \text{base}^{-2r/d}$ satisfy

$$\gamma_{\text{RoPE}} \leq 1 - \cos 1 \approx 0.4597 < \frac{N}{N-1} = \gamma^\star. \tag{4}$$

This gap motivates a frequency allocation that adapts to the available dimension budget and achieves larger γ for relative position retrieval.

4.2 Better frequency for rope in XoM

Let κ be a prime power, $\varrho \geq 2$ an integer, and $M = \kappa^\varrho - 1$. Choose a primitive element $g \in \mathbb{F}_{\kappa^\varrho}$, with subfield $\mathbb{F}_\kappa$, and define, for each $a \in \mathbb{F}_\kappa$,

$$b_a = \log_g(g + a) \pmod M, \qquad \omega_a = \frac{2\pi b_a}{M}. \tag{5}$$

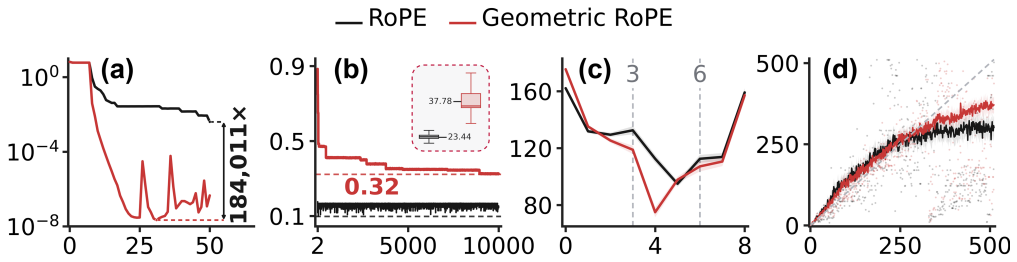

Figure 6: **Geometric RoPE improves positional retrieval.** (a) XoM test loss versus training epoch. (b) Query key margin against the strongest distractor versus marker position (main: Normalized, right: Unnoramlized). (c) A per layer MAE loss for absolute position predicting. (d) In layer 4, the predicted position x and the true position y.

Table 1: Comparable language accuracy with improved position decoding and retrieval.

(a) Language & reasoning	RoPE	**Geo. RoPE**	*(b) Grammar & structure*	RoPE	**Geo. RoPE**
PIQA (Bisk et al., 2020)	**65.72**	65.13	PAWS-Wiki (Zhang et al., 2019)	44.200	44.200
HellaSwag (Zellers et al., 2019)	29.14	**29.29**	BLiMP (Warstadt et al., 2020)	76.634	**77.415**
WinoGrande (Sakaguchi et al., 2021)	51.30	**52.88**	SyntaxGym (Gauthier et al., 2020)	71.590	**71.715**
SciQ (Welbl et al., 2017)	**68.70**	67.70	**Average (3)**	64.141	**64.443**
ARC-Easy (Clark et al., 2018)	43.33	**44.91**	*(c) Position decoding and retrieval*		
CommonsenseQA (Talmor et al., 2019)	29.57	**29.98**	**Position MAE** ↓ (Haviv et al., 2022)	94.934	**74.952**
SocialIQA (Sap et al., 2019)	42.22	**42.63**	**MQAR** (Arora et al., 2024)	2.489	**8.427**
Basic Arithmetic (AllenAI, n.d.)	23.30	**24.47**	**Position MAE:** -21.05%		
Average (8)	44.16	**44.62**	**MQAR accuracy:** $3.39\times$ **baseline**		

Accuracy (%, ↑) unless noted; position MAE is in tokens (↓).

These exponents follow the classical Bose–Chowla affine line construction (Chowla, 1962/63) and yield a normalized position code with κ rotation blocks:

$$k_j = \frac{1}{\sqrt{\kappa}} \bigoplus_{a \in \mathbb{F}_\kappa} \begin{pmatrix} \cos(j\omega_a) \\ \sin(j\omega_a) \end{pmatrix}. \tag{6}$$

We call this frequency allocation *Geometric RoPE*. Given a position dimension budget d, a deterministic default rule selects the largest power of two $\kappa \le d/2$, followed by the smallest integer $\varrho \ge 2$ such that $\kappa^\varrho - 1 \ge N$.

Applying the character sum estimate of Katz (1989) to nonzero position differences gives, for aligned queries $q_i = k_{t_i}$,

$$\gamma_{\text{Geo}} \ge 1 - \frac{\varrho - 1}{\sqrt{\kappa}}, \qquad N \le M. \tag{7}$$

Whenever $(\varrho - 1)/\sqrt{\kappa} \le \cos 1$, this certificate matches or exceeds the upper bound for standard RoPE:

$$\gamma_{\text{Geo}} \ge 1 - \frac{\varrho - 1}{\sqrt{\kappa}} \ge 1 - \cos 1 \ge \gamma_{\text{RoPE}}, \qquad N \le M. \tag{8}$$

We further prove that Geometric RoPE is asymptotically optimal when $d/N \to 0$ and $\log N/(\sqrt{d}\log d) \to 0$. Note that this is a theoretical bound under the constraints and may not be achievable in practice. Table 2 lists precomputed configurations and certified bounds for common dimension budgets. All proofs and ready-to-use frequency lists are provided in the Apppendix D.

4.3 Results of Rope with new frequency

We compare standard and Geometric RoPE on XoM using minimal models and long sequences to stress test relative position retrieval. With a single layer, a single attention head, and a sequence length of $10{,}000$, Geometric RoPE achieves substantially faster loss reduction and larger normalized and unnormalized minimum margins (Fig.6).

We further evaluate a 60M eight layer language model with Geometric RoPE at layers 3 and 6 to enhance relative position discrimination. To assess absolute position recoverability, we train a two layer probe on each layer's hidden states. As shown in the Fig.6, Geometric RoPE yields lower MAE and more accurate position estimates. Across multiple evaluation metrics, it matches or slightly exceeds the language task performance of pure RoPE while improving positional capability, supporting the transfer of geometric insights from synthetic tasks to language models.

5 Limitations

XoM isolates relative positional relationships in natural language tasks from semantic and other confounds. Whether the phenomena and explanations identified in this controlled setting fully generalize to natural language remains open. Model complexity and optimization uncertainty limit our mathematical analysis to simplified processes, precluding a complete characterization or proof of the solution reached from random initialization. Circle and simplex optima are under assumptions differ from causal XoM. Thus, they serve primarily as a plausible explanatory account, and more in depth research are required. The emphasis of this work lies in explaining the observed phenomenon and introducing a geometric analysis framework. Our language model experiments primarily assess the practical feasibility of the framework, evaluations of larger commercial models and related investigations lie beyond our primary scope and computational resources. The method's effectiveness on real world tasks still requires validation through more extensive experimentation.

6 Conclusion

We study relative position retrieval through XoM and identify a positional structure across five encodings: approximately circular, low dimensional key trajectories accompanied by target aligned marker queries. This observation connects learned positional geometry to the phase matching principle underlying rotary attention. Our analysis distinguishes the geometric usefulness of this representation from the unresolved question of why optimization selects it. Under explicit assumptions, multifrequency codes provide stronger separation than the unit circle benchmark, motivating a multifrequency alternative (Geometric Rope). Its results on language models validates the analytical method in this paper.

7 Related Work

NoPE models can recover positional information and exhibit relative attention patterns (Haviv et al., 2022; Kazemnejad et al., 2023); mean based probing also reveals low dimensional positional components in language models (Song & Zhong, 2026). We study a more specific phenomenon: shared query key rotation at a retrieval head across five non-RoPE encodings on XoM. Barbero et al. (2024) analyze position selective RoPE attention and the distinct roles of frequency bands. Position Interpolation, YaRN, and LongRoPE extend context windows by rescaling positional representations (Chen et al., 2023; Peng et al., 2024; Ding et al., 2024), while FoPE modifies their spectral structure to improve length generalization (Hua et al., 2024). Our design instead targets a worst retrieval margin over a specified address range which is distinct from a language model extrapolation guarantee. Aligned retrieval connects to spherical code bounds (Rankin, 1955). We adapts two established tools, Bose–Chowla exponents (Chowla, 1962/63) and Katz's character sum estimate (Katz, 1989), to rotary addressing; our contribution lies in the adaptation and its connection to observed positional geometry.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with translation of the manuscript. We did not use generative AI tools for any other tasks for which disclosure is required or recommended. We have reviewed all translations to ensure that they accurately preserve the intended meaning and technical content of the original text. We take full responsibility for the final content of this work, including all text translated with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed task descriptions, explanations of the theoretical analysis, and experimental settings in the main paper and the appendix. Readers can refer to these materials for the task formulations, theoretical reasoning, and experimental procedures underlying our findings. Together, these descriptions provide sufficient information to reproduce the experimental phenomena reported in the main paper.

REFERENCES

AllenAI. Basic Arithmetic dataset. Hugging Face datasets, n.d. URL `https://huggingface.co/datasets/allenai/basic_arithmetic`. Accessed September 26, 2026.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Y Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International conference on learning representations*, volume 2024, pp. 15664–15730, 2024.

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

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

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.

Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. volume 35, pp. 8386–8399, 2022.

Bose R.C. Chowla, S. Theorems in the additive theory of numbers. *Commentarii mathematici Helvetici*, 37:141–147, 1962/63.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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. 4171–4186, 2019.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. 2024.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 70–76, 2020.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, 2022.

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. 2024.

Nicholas M Katz. An estimate for character sums. *Journal of the American Mathematical Society*, 2(2):197–200, 1989.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. volume 36, pp. 24892–24928, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Robert Alexander Rankin. The closest packing of spherical caps in n dimensions. *Glasgow Mathematical Journal*, 2(3):139–144, 1955.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. volume 64, pp. 99–106. ACM New York, NY, USA, 2021.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. 2013.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018.

Jiajun Song and Yiqiao Zhong. Uncovering hidden geometry in transformers via disentangling position and context. *Harvard Data Science Review*, 8(2), 2026.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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. 4149–4158, 2019.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, et al. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. volume 30, 2017.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. Path attention: Position encoding via accumulating householder transformations. volume 38, pp. 62220–62247, 2026.

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.

Yuan Zhang, Jason Baldridge, and Luheng He. Paws: Paraphrase adversaries from word scrambling. 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. 1298–1308, 2019.

APPENDIX

A ADDRESSING OBJECTIVE, LOSS, AND THE OPTIMAL QUERY

A.1 NOTATION AND COMPARISON CLASSES

Let $\mathcal{A}$ be a finite collection of retrieval conditions, with target $t_a \in J_a \subseteq \{0, \dots, N-1\}$, $|J_a| \geq 2$, and probability $\pi_a > 0$, $\sum_a \pi_a = 1$. The physical sequence length is L; the number of addresses in a mathematical benchmark is N. Neither symbol denotes the training dataset size. Let D be the ambient real query/key dimension; the frequency-design budget d in the main text is a choice of D. We use Euclidean vector norms, Frobenius matrix norms, and operator norms throughout. Unless stated otherwise,

$$\|q_a\| \leq 1, \quad \|k_j\| \leq 1, \qquad m_{aj} = q_a^\top (k_{t_a} - k_j), \qquad \gamma = \min_{a, j \in J_a \setminus \{t_a\}} m_{aj}. \tag{A.1}$$

There is no independent additive addressing bias in this comparison class. The logit multiplier $\beta > 0$ is separate from the norm budget. In an actual head with unnormalized activations and division by $\sqrt{d_h}$, there need not be a single common multiplier after unit normalization.

We distinguish the number of addresses N in the theoretical analysis from the physical context length L. Our theorems assume $J_a = \{0, \dots, N-1\}$ for every retrieval condition and require every address to occur as a target. The causal XoM task instead has restricted candidate sets, nonuniform target coverage, and marker's legal windows. The theorems therefore provide only an explanatory benchmark.

Complete competition. The geometric optima below require $J_a = \{0, \dots, N-1\}$ and every address to occur as a target, with queries independently selectable. Causal XoM does not satisfy this assumption globally. In particular, with minimum offset two and $L = 100$, the final two addresses never occur as targets. Deleting distractors preserves a lower bound for a fixed construction, but need not preserve an upper bound or a characterization of all optima.

A.2 MEASUREMENTS AND THEIR SCOPE

To separate average positional structure from content variation, we use a mean based decomposition, closely related to the positional decomposition of Song & Zhong (2026). At a specified layer and head, let $K_{b,j}$ denote a key for content probe b at physical position j. We examine the position mean $\bar{K}_j = \mathbb{E}_b K_{b,j}$ and its centered matrix $F_K = J\bar{K}$, where $J = I - \mathbf{1}\mathbf{1}^\top / L$. For queries, $\bar{Q}_m(t)$ denotes the query observed at physical marker position $t + d_m$, indexed by the target address t. It exists only for $0 \leq t < L - d_m$, we never fill or wrap missing positions.

For a nonzero feature matrix F with singular values $\sigma_1 \geq \sigma_2 \geq \cdots$, we measure the fraction of energy captured by the leading two singular directions and their relative strength:

$$R_2(F) = \frac{\sigma_1^2 + \sigma_2^2}{\sum_r \sigma_r^2}, \qquad \zeta(F) = \frac{\sigma_2}{\sigma_1}. \tag{A.2}$$

High R_2 alone could describe a line, high ζ rules out that particular degeneracy. To test the physical frequency, define Π_r as orthogonal projection onto the cosine and sine of angular frequency $2\pi r/L$, evaluated on the actual observed positions, and set

$$e_r(F) = \frac{\|\Pi_r F\|_{\mathrm{F}}^2}{\|F\|_{\mathrm{F}}^2}. \tag{A.3}$$

A.3 WHAT CROSS-ENTROPY DOES AND DOES NOT IMPLY

The address loss of row a and its average are

$$\ell_a(\beta) = \log\left(1 + \sum_{j \neq t_a,\, j \in J_a} e^{-\beta m_{aj}}\right), \qquad R_\beta = \sum_a \pi_a \ell_a(\beta). \tag{A.4}$$

Proposition A.1 (Binary comparison and large scale address risk)**.** *For a target versus one distractor comparison, the loss is exactly* $\log(1 + e^{-\beta m_{aj}})$*. Consequently, the* worst *binary loss is* $\log(1 + e^{-\beta\gamma})$*, and minimizing it is equivalent to maximizing* γ*. This is not an identity for the average binary loss. Let* $B = \max_a(|J_a| - 1)$ *and* $\pi_0 = \min_a \pi_a$*. For* $\gamma \geq 0$*,*

$$\frac{\pi_0}{2} e^{-\beta\gamma} \leq R_\beta \leq B e^{-\beta\gamma}, \qquad \gamma - \frac{\log B}{\beta} \leq -\frac{\log R_\beta}{\beta} \leq \gamma + \frac{\log(2/\pi_0)}{\beta}. \tag{A.5}$$

Thus the exponent converges uniformly to the margin on the nonnegative margin class. On a compact representation class containing a positive margin configuration, every accumulation point of global minimizers of R_β *as* $\beta \to \infty$ *maximizes* γ*.*

Proof. For $m \in \mathbb{R}$ and $\beta > 0$,

$$-\log \frac{e^{\beta s_t}}{e^{\beta s_t} + e^{\beta s_j}} = \log(1 + e^{-\beta(s_t - s_j)}), \qquad \frac{d}{dm} \log(1 + e^{-\beta m}) = -\frac{\beta}{1 + e^{\beta m}} < 0.$$

Choose $(a_*, j_*) \in \operatorname{argmin}_{a,j} m_{aj}$. For $\gamma \geq 0$,

$$\begin{aligned} R_\beta &\leq \sum_a \pi_a \sum_{j \in J_a \setminus \{t_a\}} e^{-\beta m_{aj}} \leq \sum_a \pi_a (|J_a| - 1) e^{-\beta\gamma} \leq B e^{-\beta\gamma}, \\ R_\beta &\geq \pi_{a_*} \log(1 + e^{-\beta m_{a_* j_*}}) \geq \pi_0 \log(1 + e^{-\beta\gamma}) \geq \frac{\pi_0}{2} e^{-\beta\gamma}, \\ \gamma - \frac{\log B}{\beta} &\leq -\frac{\log R_\beta}{\beta} \leq \gamma + \frac{\log(2/\pi_0)}{\beta}. \end{aligned}$$

On the stated compact class $\mathcal{Z}$, let $z_* \in \operatorname{argmax}_{z \in \mathcal{Z}} \gamma(z)$, $\gamma^* = \gamma(z_*) > 0$, and $z_\beta \in \operatorname{argmin}_{z \in \mathcal{Z}} R_\beta(z)$.

$$\begin{aligned} \gamma(z) \leq 0 &\implies R_\beta(z) \geq \pi_0 \log 2, \\ R_\beta(z_\beta) \leq R_\beta(z_*) \leq B e^{-\beta\gamma^*} \to 0 &\implies \gamma(z_\beta) > 0, \\ \frac{\pi_0}{2} e^{-\beta\gamma(z_\beta)} \leq B e^{-\beta\gamma^*} &\implies \gamma^* - \frac{\log(2B/\pi_0)}{\beta} \leq \gamma(z_\beta) \leq \gamma^*, \\ z_{\beta_n} \to z_\infty, \quad \beta_n \to \infty &\implies \gamma(z_\infty) = \lim_n \gamma(z_{\beta_n}) = \gamma^*. \end{aligned}$$

□

Token copying is a separate statement. For an ideal pointer whose value at address j is a vocabulary one hot vector and whose readout returns the attention weighted probability vector, the probability of the correct token is

$$p(y = x_{t_a}) = \sum_{j \in J_a :\, x_j = x_{t_a}} \alpha_{aj}. \tag{A.6}$$

If no distractor carries the target token, token and address cross entropies are identical. With repeated tokens, token loss is at most address loss, and equality need not hold. In a content independent pointer with i.i.d. uniform ordinary tokens from V symbols, the expected correct token mass is $\alpha_{a,t_a} + (1 - \alpha_{a,t_a})/V$ when all distractors are ordinary tokens. This expectation is not an equality between expected log losses. Learned values, residual paths, multiple heads, and an arbitrary decoder invalidate a direct identification without additional evidence.

A.4 Convex hull characterization and alignment

For a fixed set of keys and target t, write

$$C_t = \operatorname{conv}\{k_j : j \neq t\}, \quad p_t = \operatorname*{argmin}_{x \in C_t} \|k_t - x\|, \quad h_t = \|k_t - p_t\|.$$

For incomplete competition, use the row's distractor hull instead.

Theorem A.2 (Optimal query and exact alignment criterion)**.** *For arbitrary fixed keys,*

$$\max_{\|q\|\le 1} \min_{j\neq t} q^\top (k_t - k_j) = h_t. \tag{A.7}$$

If $h_t > 0$, the unique maximizer is $(k_t - p_t)/h_t$; if $h_t = 0$, the zero query attains zero. For distinct unit keys let $\mu_t = \max_{j\neq t} k_t^\top k_j$ and $I_t = \{j \neq t : k_t^\top k_j = \mu_t\}$. Then $h_t > 0$ and

$$q_t^\star = k_t \quad \Longleftrightarrow \quad \mu_t k_t \in \operatorname{conv}\{k_j : j \in I_t\}. \tag{A.8}$$

In that case $h_t = 1 - \mu_t$ and the aligned margin is half the nearest squared key distance. Under complete competition, optimizing all queries gives $\min_t h_t$. If all h_t are equal and positive, every query at a global maximizer is individually optimal.

Proof. Compactness gives a minimizer $p_t \in C_t$. For $p, p' \in C_t$,

$$\left\| k_t - \frac{p+p'}{2} \right\|^2 = \frac{\|k_t - p\|^2 + \|k_t - p'\|^2}{2} - \frac{\|p - p'\|^2}{4},$$

so distinct minimizers are impossible. For $x \in C_t$,

$$0 \le \frac{d}{ds}\|k_t - p_t - s(x - p_t)\|^2\Big|_{s=0^+} = -2(k_t - p_t)^\top (x - p_t),$$
$$\min_{j\neq t} q^\top (k_t - k_j) = \min_{x\in C_t} q^\top (k_t - x) \le q^\top (k_t - p_t) \le \|q\| h_t \le h_t.$$

If $h_t > 0$, $q_* = (k_t - p_t)/h_t$ satisfies

$$q_*^\top (k_t - x) = h_t - \frac{(k_t - p_t)^\top (x - p_t)}{h_t} \ge h_t.$$

Equality in Cauchy–Schwarz forces $q = q_*$; if $h_t = 0$, $q = 0$ attains zero. For distinct unit keys, $\mu_t < 1$ and

$$\max_{x\in C_t} k_t^\top x = \mu_t < 1 = k_t^\top k_t \implies k_t \notin C_t \implies h_t > 0.$$

For $p = \mu_t k_t \in \operatorname{conv}\{k_j : j \in I_t\}$,

$$(k_t - p)^\top (k_j - p) = (1 - \mu_t)(k_t^\top k_j - \mu_t) \le 0 \implies p = p_t, \quad h_t = 1 - \mu_t, \quad q_* = k_t.$$

Conversely,

$$q_* = k_t \implies h_t = \min_{j\neq t} k_t^\top (k_t - k_j) = 1 - \mu_t, \quad p_t = k_t - h_t k_t = \mu_t k_t,$$
$$p_t = \sum_{j\neq t} \lambda_j k_j, \quad \lambda_j \ge 0, \quad \sum_j \lambda_j = 1$$
$$\implies \sum_{j\neq t} \lambda_j (\mu_t - k_t^\top k_j) = 0 \implies (\lambda_j > 0 \Rightarrow j \in I_t).$$

Finally,

$$1 - k_t^\top k_j = \tfrac{1}{2}\|k_t - k_j\|^2, \qquad \max_{\{\|q_a\|\le 1\}} \gamma = \min_t h_t.$$

If all $h_t = h > 0$, attaining $\gamma = h$ forces each row to attain its unique maximizer above. □

B Optimal geometry and the limitation of a single frequency

B.1 The simplex benchmark in sufficient dimension

Theorem B.1 (Free high dimensional optimum)**.** *For $N \ge 3$ and $D \ge N - 1$ under complete competition,*

$$\gamma^\star = \frac{N}{N-1}. \tag{B.1}$$

All maximizing keys form a unit regular simplex, up to orthogonal transformation and relabeling, and every query equals its target key.

Proof. Choose one query u_t per target and set $\bar{k} = N^{-1}\sum_t k_t$.

$$N(N-1)\gamma \le \sum_t \sum_{j\neq t} u_t^\top (k_t - k_j) = N\sum_t u_t^\top (k_t - \bar{k})$$
$$\le N\Big(\sum_t \|u_t\|^2\Big)^{1/2}\Big(\sum_t \|k_t\|^2 - N\|\bar{k}\|^2\Big)^{1/2} \le N^2.$$

For $k_t = \sqrt{N/(N-1)}(e_t - \mathbf{1}/N)$ and $u_t = k_t$,

$$\mathbf{1}^\top k_t = 0, \quad \|k_t\| = 1, \quad k_t^\top k_j = -\frac{1}{N-1}\ (t \neq j), \quad \gamma = 1 + \frac{1}{N-1} = \frac{N}{N-1}.$$

Equality throughout the upper-bound chain implies

$$\|k_t\| = \|u_t\| = 1, \quad \bar{k} = 0, \quad u_t = k_t, \quad u_t^\top (k_t - k_j) = \frac{N}{N-1} \quad (t \neq j).$$

Thus the Gram matrix is $N(I - \mathbf{1}\mathbf{1}^\top/N)/(N-1)$, uniquely determining the simplex up to an isometry. For every additional query with target t,

$$\frac{N}{N-1} \le \frac{1}{N-1}\sum_{j\neq t} q_a^\top (k_t - k_j) = \frac{N}{N-1} q_a^\top k_t \le \frac{N}{N-1} \implies q_a = k_t.$$

□

B.2 Unit circle optimum and the role of physical order

Theorem B.2 (Circle optimum with free queries)**.** *Let $N \ge 3$, require keys to lie on the* unit circle *in $\mathbb{R}^2$, and allow queries in the closed unit disk. Under complete competition,*

$$\gamma_\circ^\star = 1 - \cos(2\pi/N). \tag{B.2}$$

The maximizing unordered key set is precisely a regular N-gon, with target aligned queries. Among these maximizers, minimizing $E_{\text{path}} = \sum_{j=0}^{N-2} \|k_{j+1} - k_j\|^2$ forces

$$k_j = (\cos(\phi + \omega j), \sin(\phi + \omega j))^\top, \quad \omega = \pm 2\pi/N, \quad E_{\text{path}} = 2(N-1)\gamma_\circ^\star. \tag{B.3}$$

Proof. Coincident keys give $\gamma \le 0$. Otherwise, order the keys cyclically with gaps $\Delta_i > 0$ and $\sum_i \Delta_i = 2\pi$ (indices modulo N). For neighboring gaps $a, b \in (0, \pi)$, rotate the middle key to $(1, 0)$. The neighboring chord has normal $n = (\cos((b-a)/2), \sin((b-a)/2))$ and equation $n^\top x = \cos((a+b)/2)$. Its perpendicular foot is on the segment since

$$\sin^2((a+b)/2) - \sin^2((b-a)/2) = \sin a \sin b > 0.$$

Therefore

$$h_i \le \cos((b-a)/2) - \cos((a+b)/2) = 2\sin(a/2)\sin(b/2) \le 2\sin^2((a+b)/4). \tag{B.4}$$

If one gap is at least π, write the remaining gaps as $\delta_1, \ldots, \delta_{N-1} > 0$, with $S = \sum_r \delta_r \le \pi$. For $N = 3$, $h_i \le 2\sin^2(S/4) \le 1 < 3/2$ at the middle vertex. For $N \ge 4$,

$$\sum_{r=1}^{N-2} (\delta_r + \delta_{r+1}) = 2S - \delta_1 - \delta_{N-1} < 2\pi,$$
$$\exists r: \quad \delta_r + \delta_{r+1} < \frac{2\pi}{N-2},$$
$$\min_i h_i < 2\sin^2\left(\frac{\pi}{2(N-2)}\right) \le 2\sin^2(\pi/N).$$

Otherwise $0 < \Delta_i < \pi$ for all i, and

$$\min_i h_i \le \left(\prod_i 2\sin(\Delta_{i-1}/2)\sin(\Delta_i/2)\right)^{1/N} = 2\left(\prod_i \sin(\Delta_i/2)\right)^{2/N},$$

$$\frac{d^2}{dx^2}\log\sin(x/2) = -\tfrac{1}{4}\csc^2(x/2) < 0,$$

$$\frac{1}{N}\sum_i \log\sin(\Delta_i/2) \le \log\sin(\pi/N),$$

$$\min_i h_i \le 2\sin^2(\pi/N) = 1 - \cos(2\pi/N).$$

Equality requires $\Delta_i = 2\pi/N$ for every i. The regular polygon attains the bound, with

$$\tfrac{1}{2}(k_{i-1} + k_{i+1}) = \cos(2\pi/N)k_i \in C_i, \qquad q_i^* = k_i$$

by Theorem A.2. Let $\ell_j \in \mathbb{Z}/N\mathbb{Z}$ label the vertex at physical address j. Then

$$\|k_{j+1} - k_j\|^2 \ge 2\gamma_\circ^*, \qquad E_{\text{path}} \ge 2(N-1)\gamma_\circ^*,$$
$$E_{\text{path}} = 2(N-1)\gamma_\circ^* \implies \ell_{j+1} - \ell_j \equiv \epsilon_j \pmod N, \quad \epsilon_j \in \{-1, 1\},$$
$$\epsilon_{j+1} = -\epsilon_j \implies \ell_{j+2} = \ell_j \quad (\perp),$$
$$\ell_j = \ell_0 + j\epsilon \pmod N, \quad \epsilon \in \{-1, 1\}.$$

This is precisely (B.3). □

Remark B.3 (Scope of the circular explanation). The theorem uses unit circle keys, not arbitrary points in the unit disk. It first maximizes the margin and then minimizes path energy within that optimal set. At $N = 100$, the benchmark is 0.0019732716, and it scales as $2\pi^2/N^2 + O(N^{-4})$.

B.3 When a shared orthogonal step exists

Proposition B.4 (Cyclic Gram matrix and orthogonal transport)**.** *For a finite cyclic family $k_0, \dots, k_{N-1}$, an orthogonal map on their span with $Uk_j = k_{j+1 \bmod N}$ exists if and only if $k_{i+1}^\top k_{j+1} = k_i^\top k_j$ for all i, j. It is unique on the span and satisfies $U^N = I$ there.*

Proof. Orthogonal transport implies $k_{i+1}^\top k_{j+1} = k_i^\top U^\top U k_j = k_i^\top k_j$. Conversely, put $K = [k_0, \dots, k_{N-1}]$ and $K_+ = [k_1, \dots, k_0]$. Gram invariance gives

$$K_+^\top K_+ = K^\top K, \qquad \|K_+ c\|^2 = c^\top K^\top K c = \|Kc\|^2.$$

Hence $Kc = 0 \Rightarrow K_+ c = 0$, and $U(Kc) := K_+ c$ is well-defined. For all c, d,

$$\langle U(Kc), U(Kd)\rangle = c^\top K_+^\top K_+ d = \langle Kc, Kd\rangle, \quad \operatorname{range} U = \operatorname{span}\{k_j\}, \quad U^N k_j = k_j.$$

The spanning family determines U uniquely and yields $U^N = I$ on its span. □

Every real orthogonal operator decomposes into plane rotations and scalar $+1$ or -1 blocks. Repeating the same nonzero frequency across many coordinate planes does not create many independent positional modes:

$$R(j\omega)u \text{ in all such blocks} \quad \implies \quad k_j = a\cos(j\omega) + b\sin(j\omega),$$

so the positional matrix has rank at most two.

Full spectrum realization of the simplex. For odd N, the construction in (3) has kernel

$$k_i^\top k_j = \frac{2}{N-1}\sum_{r=1}^{(N-1)/2} \cos(2\pi r(i-j)/N) = -\frac{1}{N-1} \quad (i \ne j),$$

because the sum of all Nth roots of unity is zero. For even N, use rotation blocks $r = 1, \dots, N/2-1$ with weight $2/(N-1)$ and one scalar Nyquist coordinate $(-1)^j/\sqrt{N-1}$. Its kernel is

$$\frac{2\sum_{r=1}^{N/2-1}\cos(2\pi r\Delta/N) + (-1)^\Delta}{N-1}.$$

This again equals $-1/(N-1)$ for nonzero residues.

B.4 A STRICT IMPROVEMENT USING TWO FREQUENCIES

Proposition B.5 (Rank four improvement). *For $N \geq 5$ and $D \geq 4$, there is a two frequency aligned code whose margin strictly exceeds (B.2).*

Proof. Put $\theta = 2\pi/N$, $A_r(j) = 1 - \cos(rj\theta)$, and

$$g_1 = 1 - \cos\theta, \qquad \delta = \min_{2 \leq j \leq N-2} (A_1(j) - g_1) > 0, \qquad b = \cos\theta - \cos 2\theta > 0.$$

For $0 < \epsilon < \min\{1, \delta/4\}$, define

$$k_j = \sqrt{1-\epsilon}(\cos j\theta, \sin j\theta) \oplus \sqrt{\epsilon}(\cos 2j\theta, \sin 2j\theta).$$

Its aligned gaps satisfy

$$\begin{aligned} 1 - k_0^\top k_j &= (1-\epsilon)A_1(j) + \epsilon A_2(j), \\ j = \pm 1 \pmod N &\implies 1 - k_0^\top k_j = g_1 + \epsilon b > g_1, \\ 2 \leq j \leq N-2 &\implies 1 - k_0^\top k_j \geq g_1 + \delta - 2\epsilon > g_1. \end{aligned}$$

For $N \geq 5$, the residues $1, -1, 2, -2$ are distinct modulo N and

$$\sum_{j=0}^{N-1} e^{2\pi \mathrm{i}(r-s)j/N} = N\mathbf{1}_{r=s} \qquad (r, s \in \{1, -1, 2, -2\}),$$

so the four real Fourier columns are independent and the code has rank four. □

This establishes strict improvement over the specified circular benchmark.

C POSSIBLE MECHANISM FOR LOW DIMENSIONAL LEARNING

C.1 AN EXACT BACKPROPAGATION IDENTITY

For sample or accumulated microbatch b at step n, write

$$\begin{aligned} Q^{(b)} &= X_Q^{(b)} U, \qquad K^{(b)} = X_K^{(b)} V, \qquad S^{(b)} = \beta Q^{(b)} K^{(b)\top} + B^{(b)}, \\ G^{(b)} &= \frac{\partial \mathcal{L}_n}{\partial S^{(b)}}, \qquad C_n = -\beta \sum_b X_Q^{(b)\top} G^{(b)} X_K^{(b)}. \end{aligned} \tag{C.1}$$

Here $U \in \mathbb{R}^{p_Q \times d_h}$ and $V \in \mathbb{R}^{p_K \times d_h}$ are the query and *key* projections; V is not the value projection. Batch averaging is already included in G. Masked entries have zero gradient. $\mathcal{L}_n$ may be the actual token loss. The identity below requires independent projection parameters, no other path from these same parameters to the loss, and a dot product after projection. Features and additive bias must not depend on the same U, V through another path. Changing upstream features across steps is allowed.

Lemma C.1 (bilinear drive). *Under these architecture assumptions, excluding separately handled decay,*

$$\nabla_U \mathcal{L}_n = -C_n V_n, \qquad \nabla_V \mathcal{L}_n = -C_n^\top U_n. \tag{C.2}$$

With $W_n = (U_n^\top, V_n^\top)^\top$ and $A_n = \left(\begin{smallmatrix} 0 & C_n \\ C_n^\top & 0 \end{smallmatrix}\right)$, any observed update has the exact decomposition

$$W_{n+1} = \big[(1 - \eta_n \lambda_n) I + \eta_n A_n\big] W_n + R_n, \tag{C.3}$$

where R_n is defined by subtraction. Plain SGD with common decay has $R_n = 0$; AdamW, momentum, and clipping generally do not.

Proof. For a perturbation D_U and the Frobenius pairing,

$$\begin{aligned}
D_U S^{(b)} &= \beta X_Q^{(b)} D_U V^\top X_K^{(b)\top}, \\
D_U \mathcal{L}_n &= \beta \sum_b \mathrm{tr}\Big(G^{(b)\top} X_Q^{(b)} D_U V^\top X_K^{(b)\top}\Big) \\
&= \left\langle \beta \sum_b X_Q^{(b)\top} G^{(b)} X_K^{(b)} V, D_U \right\rangle_F = \langle -C_n V, D_U \rangle_F, \\
D_V \mathcal{L}_n &= \left\langle \beta \sum_b X_K^{(b)\top} G^{(b)\top} X_Q^{(b)} U, D_V \right\rangle_F = \langle -C_n^\top U, D_V \rangle_F, \\
\nabla_W \mathcal{L}_n &= -A_n W_n, \qquad R_n := W_{n+1} - [(1-\eta_n\lambda_n)I + \eta_n A_n] W_n.
\end{aligned}$$

□

C.2 Two direction amplification

Let $C = \sum_r s_r \ell_r r_r^\top$ be its compact SVD, with $s_1 \geq s_2 \geq s_3 \geq 0$, padding missing singular values with zero. Define

$$v_r^+ = 2^{-1/2}(\ell_r^\top, r_r^\top)^\top, \qquad v_r^- = 2^{-1/2}(\ell_r^\top, -r_r^\top)^\top, \qquad P = v_1^+ v_1^{+\top} + v_2^+ v_2^{+\top}.$$

These are directions in the input feature spaces of the projections.

Theorem C.2 (Local concentration with optimizer residual)**.** *Suppose $s_2 > s_3$, $\lambda \geq 0$, and $0 < \eta < 1/(\lambda + s_1)$. For the current-step projector P, let*

$$\begin{aligned}
&a = \|PW\|_F, \qquad b = \|(I-P)W\|_F, \qquad e_a = \|PR\|_F, \qquad e_b = \|(I-P)R\|_F, \\
&\alpha = 1 - \eta\lambda + \eta s_2, \qquad \nu = 1 - \eta\lambda + \eta s_3.
\end{aligned}$$

Writing $a^+ = \|PW^+\|_F$, $b^+ = \|(I-P)W^+\|_F$, one has

$$a^+ \geq \alpha a - e_a, \qquad b^+ \leq \nu b + e_b, \qquad 0 < \nu < \alpha. \tag{C.4}$$

If $a, b > 0$ and $\alpha a > e_a$, then

$$\frac{b^+}{a^+} \leq \frac{\nu b + e_b}{\alpha a - e_a}. \quad \textit{In particular,} \quad \eta(s_2 - s_3)ab > a e_b + b e_a \implies \frac{b^+}{a^+} < \frac{b}{a}. \tag{C.5}$$

The gain of mode $v_r^\pm$ is exactly

$$v_r^{\pm\top} W^+ = (1 - \eta\lambda \pm \eta s_r) v_r^{\pm\top} W + v_r^{\pm\top} R. \tag{C.6}$$

Proof. Let $T = (1-\eta\lambda)I + \eta A$. The SVD identities give

$$\begin{aligned}
A v_r^\pm = \pm s_r v_r^\pm, \qquad T v_r^\pm = (1 - \eta\lambda \pm \eta s_r) v_r^\pm, \\
\lambda_{\min}(T) \geq 1 - \eta(\lambda + s_1) > 0, \quad PT = TP, \\
\lambda_{\min}(T|_{\mathrm{range}\, P}) = \alpha, \qquad \|T|_{\mathrm{range}(I-P)}\|_{\mathrm{op}} \leq \nu < \alpha, \\
a^+ = \|TPW + PR\|_F \geq \|TPW\|_F - \|PR\|_F \geq \alpha a - e_a, \\
b^+ = \|T(I-P)W + (I-P)R\|_F \leq \nu b + e_b.
\end{aligned}$$

For $a, b > 0$ and $\alpha a > e_a$,

$$\begin{aligned}
\frac{b^+}{a^+} &\leq \frac{\nu b + e_b}{\alpha a - e_a}, \\
\frac{\nu b + e_b}{\alpha a - e_a} < \frac{b}{a} &\iff a(\nu b + e_b) < b(\alpha a - e_a) \\
&\iff a e_b + b e_a < (\alpha - \nu) ab = \eta(s_2 - s_3) ab, \\
v_r^{\pm\top} W^+ &= v_r^{\pm\top}(TW + R) = (1 - \eta\lambda \pm \eta s_r) v_r^{\pm\top} W + v_r^{\pm\top} R.
\end{aligned}$$

□

Moving directions cannot be ignored. If the next drive defines P^+, set $d_P = \|P^+ - P\|_{\mathrm{op}}\|W^+\|_F$. Then

$$\|P^+W^+\|_F \geq \alpha a - e_a - d_P, \qquad \|(I-P^+)W^+\|_F \leq \nu b + e_b + d_P.$$

These follow by adding and subtracting PW^+.

C.3 Reading parameter concentration out as positional geometry

Let $\bar{X}_K$ be the centered mean hidden feature matrix on a fixed positional probe. Define $F = \bar{X}_K V$, $R_2^{\mathrm{in}} = [r_1, r_2]$, $Z = [v_1^+, v_2^+]^\top W$, and

$$F_2 = 2^{-1/2}\bar{X}_K R_2^{\mathrm{in}} Z, \qquad E = F - F_2.$$

Proposition C.3 (Rank, visibility, and frequency readout)**.** *The following deterministic bounds hold:*

$$\sum_{j\geq 3}\sigma_j(F)^2 \leq \|E\|_F^2 \leq \|\bar{X}_K\|_{\mathrm{op}}^2 b^2, \tag{C.7}$$

$$\sigma_2(F) \geq \frac{\sigma_2(\bar{X}_K R_2^{\mathrm{in}})\sigma_2(Z)}{\sqrt{2}} - \|E\|_F. \tag{C.8}$$

For any fixed orthogonal physical frequency projector Π_1 and $F \neq 0$,

$$1 - \frac{\|\Pi_1 F\|_F^2}{\|F\|_F^2} \leq \frac{\left(\|(I-\Pi_1)\bar{X}_K R_2^{\mathrm{in}}\|_{\mathrm{op}}\|Z\|_F/\sqrt{2} + \|E\|_F\right)^2}{\|F\|_F^2}. \tag{C.9}$$

Proof. Let S_K select the key block of a stacked projection. Then

$$\begin{aligned}
S_K PW = R_2^{\mathrm{in}} Z/\sqrt{2}, \quad V - S_K PW = S_K(I-P)W, \quad \|S_K\|_{\mathrm{op}} = 1,\\
E = \bar{X}_K S_K (I-P)W, \quad \|E\|_F \leq \|\bar{X}_K\|_{\mathrm{op}}\|(I-P)W\|_F = \|\bar{X}_K\|_{\mathrm{op}} b,\\
\operatorname{rank} F_2 \leq 2 \implies \sum_{j\geq 3}\sigma_j(F)^2 = \min_{\operatorname{rank} H \leq 2}\|F-H\|_F^2 \leq \|F-F_2\|_F^2 = \|E\|_F^2,
\end{aligned}$$

$$\begin{aligned}
\sigma_2(F) \geq \sigma_2(F_2) - \|E\|_{\mathrm{op}} \geq \frac{\sigma_2(\bar{X}_K R_2^{\mathrm{in}})\sigma_2(Z)}{\sqrt{2}} - \|E\|_F,\\
\|(I-\Pi_1)F\|_F \leq \|(I-\Pi_1)F_2\|_F + \|(I-\Pi_1)E\|_F\\
\leq \frac{\|(I-\Pi_1)\bar{X}_K R_2^{\mathrm{in}}\|_{\mathrm{op}}\|Z\|_F}{\sqrt{2}} + \|E\|_F,\\
1 - \frac{\|\Pi_1 F\|_F^2}{\|F\|_F^2} = \frac{\|(I-\Pi_1)F\|_F^2}{\|F\|_F^2}.
\end{aligned}$$

□

The right side of (C.8) must be positive to rule out rank one. Frequency concentration additionally requires the measured spatial shape in (C.9).

C.4 Why a two dimensional stage is plausible but not inevitable

Proposition C.4 (Minimal separable rank)**.** *For $N \geq 3$ under complete competition and without additive addressing bias, rank one keys cannot have positive margin for every target. Rank two suffices.*

Proof. Write $k_j = c_j u$ with $c_1 \leq \cdots \leq c_N$. If $c_1 < c_N$, then

$$k_2 = \frac{c_N - c_2}{c_N - c_1}k_1 + \frac{c_2 - c_1}{c_N - c_1}k_N \in C_2;$$

if $c_1 = c_N$, all keys coincide. In either case,

$$\min_t \operatorname{dist}(k_t, C_t) = 0.$$

Theorem A.2 rules out positive complete-competition margin. A regular N-gon has rank two and margin $1 - \cos(2\pi/N) > 0$. □

A model of sequential mode growth. Consider $\mathcal{L}(U,V) = \|UV^\top - Y\|_F^2/2$ with $Y = \mathrm{diag}(y_1,\dots,y_p)$, $y_r > 0$, and balanced diagonal initialization $U(0) = V(0) = \epsilon I$. Gradient flow preserves diagonality and balance. Writing the squared mode amplitude as z_r, direct differentiation gives

$$\dot{z}_r = 2z_r(y_r - z_r), \qquad z_r(t) = \frac{y_r}{1 + (y_r/\epsilon^2 - 1)e^{-2y_r t}}. \tag{C.10}$$

For $\epsilon^2 < y_r/2$, the positive half saturation time is $[2y_r]^{-1}\log(y_r/\epsilon^2 - 1)$. Separated second and third activation times can create an intermediate two mode regime, although the target is full rank and all modes eventually grow. This calculation illustrates mode dependent learning (Saxe et al., 2013); it does not identify the Transformers' unknown effective targets with this diagonal model.

A finite remaining drive limits tail regrowth. For fixed P, decompose $W = X + Y$ with $X = PW$, $Y = (I-P)W$. In a continuous update $\dot{W} = A(t)W + E(t)$, let the largest eigenvalue of the symmetric part of $(I-P)A(t)(I-P)$ be at most $b(t)$ and let $f(t) = \|(I-P)A(t)PW\|_F + \|(I-P)E(t)\|_F$. The norm derivative and an integrating factor give

$$\|Y(t)\|_F \le e^{\int_s^t b(u)du}\|Y(s)\|_F + \int_s^t e^{\int_v^t b(u)du} f(v)\,dv. \tag{C.11}$$

Consequently, a weak tail stays weak over a finite window if the residual, cross direction input, and integrated growth remain small.

C.5 From ordered circle points to a single physical frequency

Proposition C.5 (Conditional equalization of angular gaps)**.** *Let $k_i = (\cos\phi_i, \sin\phi_i)$ have cyclically ordered positive gaps $\Delta_i = \phi_{i+1} - \phi_i$ with total 2π, and let $\theta = 2\pi/N$. For the explicit surrogate*

$$\mathcal{E}(\phi) = \sum_i f(\Delta_i), \qquad f(x) = 2e^{-\beta(1-\cos x)},$$

suppose all initial gaps lie in $[a,b] \subset (0,\pi)$ containing θ and $m = \min_{[a,b]} f'' > 0$. Along its Euclidean gradient flow,

$$\dot{\Delta}_i = f'(\Delta_{i+1}) - 2f'(\Delta_i) + f'(\Delta_{i-1}),$$
$$\mathcal{V}(t) := \sum_i (\Delta_i - \theta)^2 \le \mathcal{V}(0)e^{-8m\sin^2(\pi/N)t}. \tag{C.12}$$

If Π_1 projects onto the full grid sine/cosine pair at $2\pi/N$,

$$1 - \frac{\|\Pi_1 K(t)\|_F^2}{\|K(t)\|_F^2} \le \frac{\mathcal{V}(t)}{4N\sin^2(\pi/N)} \longrightarrow 0. \tag{C.13}$$

Proof. With cyclic indices,

$$\dot{\phi}_i = -\partial_{\phi_i}\mathcal{E} = f'(\Delta_i) - f'(\Delta_{i-1}),$$
$$\dot{\Delta}_i = \dot{\phi}_{i+1} - \dot{\phi}_i = f'(\Delta_{i+1}) - 2f'(\Delta_i) + f'(\Delta_{i-1}),$$
$$f''(x) = 2\beta e^{-\beta(1-\cos x)}(\beta\sin^2 x - \cos x) \ge m > 0,$$
$$i \in \mathrm{argmax}_j\,\Delta_j \implies \dot{\Delta}_i \le 0, \qquad i \in \mathrm{argmin}_j\,\Delta_j \implies \dot{\Delta}_i \ge 0,$$
$$\Delta(0) \in [a,b]^N \implies \Delta(t) \in [a,b]^N, \qquad \sum_i \Delta_i(t) = 2\pi.$$

Summation by parts and $f'' \ge m$ give

$$\tfrac{1}{2}\dot{\mathcal{V}} = -\sum_i (\Delta_{i+1} - \Delta_i)(f'(\Delta_{i+1}) - f'(\Delta_i))$$
$$= -\sum_i (\Delta_{i+1} - \Delta_i)^2 \int_0^1 f''(\Delta_i + s(\Delta_{i+1} - \Delta_i))\,ds$$
$$\le -m\sum_i (\Delta_{i+1} - \Delta_i)^2.$$

For $(Dz)_i = z_{i+1} - z_i$, its Fourier eigenvalues are

$$D^\top D: \quad 4\sin^2(\pi r/N), \quad 0 \le r < N; \qquad \mathbf{1}^\top z = 0 \implies \|Dz\|^2 \ge 4\sin^2(\pi/N)\|z\|^2.$$

Consequently,

$$\dot{\mathcal{V}} \le -8m\sin^2(\pi/N)\mathcal{V} \implies \mathcal{V}(t) \le \mathcal{V}(0)e^{-8m\sin^2(\pi/N)t}.$$

Set $u_i = \phi_i - i\theta - \bar{\phi}$ with $\sum_i u_i = 0$, and $K_i^* = (\cos(i\theta + \bar{\phi}), \sin(i\theta + \bar{\phi}))$. Then

$$(Du)_i = \Delta_i - \theta, \quad \sum_i u_i^2 \le \frac{\mathcal{V}}{4\sin^2(\pi/N)}, \quad \Pi_1 K^* = K^*,$$

$$\|K - K^*\|_F^2 = \sum_i 4\sin^2(u_i/2) \le \sum_i u_i^2,$$

$$1 - \frac{\|\Pi_1 K\|_F^2}{\|K\|_F^2} = \frac{\|(I - \Pi_1)K\|_F^2}{N} \le \frac{\|K - K^*\|_F^2}{N} \le \frac{\mathcal{V}(t)}{4N\sin^2(\pi/N)} \to 0.$$

□

Together with the rank four improvement of Proposition B.5, it clarifies why a circular representation can coexist with a strictly better higher dimensional code.

D GEOMETRIC ROPE

D.1 A COMPARISON THAT ALLOWS FREE QUERIES

Let $\Gamma_{N,d}$ be the optimal margin over all unit ball queries and keys in $\mathbb{R}^d$ under complete competition. For even d, define the standard fixed class by

$$k_j = \bigoplus_{r=0}^{d/2-1} R(j\omega_r)u_r, \quad \sum_r \|u_r\|^2 \le 1, \quad \omega_r = B^{-2r/d}, \quad B > 1,$$

while allowing each query to be chosen freely in the unit ball. Denote its best margin by $\Gamma_{N,d}^{\mathrm{std}}$.

Theorem D.1 (Local midpoint obstruction)**.** *Suppose some row has target t and both $t-1, t+1$ as distractors. For a fixed orthogonal orbit with all principal angular frequencies bounded by $\omega_{\max} \le \pi$, every unit query satisfies*

$$\gamma \le \left[\sum_r w_r(1 - \cos\omega_r)^2\right]^{1/2} \le 1 - \cos\omega_{\max}, \quad w_r = \|u_r\|^2. \tag{D.1}$$

In particular, $\Gamma_{N,d}^{\mathrm{std}} \le c_{\mathrm{std}} = 1 - \cos 1 = 0.4596976941\ldots$ for every complete contiguous window $N \ge 3$.

Proof. Let $m_\pm = q^\top(k_t - k_{t\pm1})$ and $\bar{k} = (k_{t-1} + k_{t+1})/2$.

$$\gamma \le \min(m_-, m_+) \le \tfrac{1}{2}(m_- + m_+) = q^\top(k_t - \bar{k})$$
$$\le \|q\|\|k_t - \bar{k}\| \le \|k_t - \bar{k}\|,$$
$$\tfrac{1}{2}\big(R(-\omega_r) + R(\omega_r)\big) = \cos\omega_r I_2,$$
$$k_t - \bar{k} = \bigoplus_r R(t\omega_r)(1 - \cos\omega_r)u_r,$$
$$\|k_t - \bar{k}\|^2 = \sum_r \|u_r\|^2(1 - \cos\omega_r)^2 \le (1 - \cos\omega_{\max})^2 \sum_r w_r \le (1 - \cos\omega_{\max})^2,$$
$$\omega_r = B^{-2r/d} \le 1 \le \pi, \quad N \ge 3 \implies \Gamma_{N,d}^{\mathrm{std}} \le 1 - \cos 1.$$

□

No key and query alignment assumption is used in this upper bound.

D.2 FREQUENCY ALLOCATION

Use the main text notation throughout:

$$\kappa = p^s \ (p \text{ prime},\ s \geq 1), \qquad \varrho \in \mathbb{Z}_{\geq 2}, \qquad F = \mathbb{F}_\kappa \subset E = \mathbb{F}_{\kappa^\varrho}, \qquad M = \kappa^\varrho - 1, \quad E^\times = \langle g \rangle.$$

The Bose–Chowla allocation (Chowla, 1962/63) is

$$b_a = \log_g(g + a) \pmod M, \quad 1 \leq b_a \leq M - 1, \qquad \omega_a = \frac{2\pi b_a}{M}, \quad a \in F. \tag{D.2}$$

Well definedness, distinctness, and exact period.

$$\begin{aligned}
\nu := [F(g) : F] \mid \varrho, \qquad g \in \mathbb{F}_{\kappa^\nu}^\times \implies M = \operatorname{ord}(g) \mid \kappa^\nu - 1,\\
\nu < \varrho \implies M \leq \kappa^\nu - 1 < \kappa^\varrho - 1 = M \quad (\perp),\\
[F(g) : F] = \varrho, \qquad F(g) = E, \qquad g \notin F,\\
g + a = 0 \implies g = -a \in F \quad (\perp),\\
g + a = 1 \implies g = 1 - a \in F \quad (\perp),\\
g + a \in E^\times \setminus \{1\} \implies \exists!\, b_a \in \{1, \dots, M - 1\} : g^{b_a} = g + a,\\
b_a = b_{a'} \implies g + a = g^{b_a} = g^{b_{a'}} = g + a' \implies a = a',\\
g^{b_0} = g \implies b_0 \equiv 1 \pmod M \implies b_0 = 1.
\end{aligned}$$

Define the active code and its one step rotation by

$$v_j = \frac{1}{\sqrt{\kappa}} \bigoplus_{a \in F} \begin{pmatrix} \cos(j\omega_a) \\ \sin(j\omega_a) \end{pmatrix}, \qquad U = \bigoplus_{a \in F} R(\omega_a). \tag{D.3}$$

$$\begin{aligned}
\|v_j\|^2 = \frac{1}{\kappa} \sum_{a \in F} \big(\cos^2(j\omega_a) + \sin^2(j\omega_a)\big) = \frac{|F|}{\kappa} = 1,\\
U^\top U = I_{2\kappa}, \qquad v_{j+1} = U v_j, \qquad v_j \in \mathbb{R}^{2\kappa},\\
v_{j+T} = v_j \iff e^{\mathrm{i} T \omega_a} = 1 \quad (\forall a \in F)\\
\iff M \mid T b_a \quad (\forall a \in F) \iff M \mid T,\\
\min\{T \in \mathbb{Z}_{>0} : v_{j+T} = v_j \ \forall j\} = M.
\end{aligned}$$

Character sum input (Katz, Theorem 1). The only external estimate used below is Katz (1989):

$$\left.\begin{aligned} &|F| = \kappa, \quad [E : F] = \varrho, \quad F(x) = E,\\ &\chi : E^\times \to \{z \in \mathbb{C} : |z| = 1\}, \quad \chi(yz) = \chi(y)\chi(z), \quad \chi \not\equiv 1 \end{aligned}\right\} \implies \left| \sum_{a \in F} \chi(x + a) \right| \leq (\varrho - 1)\sqrt{\kappa}. \tag{D.4}$$

The quadratic case is proved independently in Section D.3.

Theorem D.2 (All offset coherence and margin certificate)**.** *Let $\varepsilon = (\varrho - 1)/\sqrt{\kappa}$. Then*

$$\Delta \not\equiv 0 \pmod M \implies \left| \frac{1}{\kappa} \sum_{a \in F} e^{2\pi \mathrm{i} \Delta b_a / M} \right| \leq \varepsilon, \qquad |v_i^\top v_j| \leq \varepsilon \quad (i - j \not\equiv 0 \pmod M). \tag{D.5}$$

For addresses distinct modulo M, keys $k_j = v_j$ and aligned queries $q_a = v_{t_a}$ satisfy

$$\gamma_{\mathrm{Geo}} = \min_a \min_{j \in J_a \setminus \{t_a\}} q_a^\top (k_{t_a} - k_j) \geq 1 - \varepsilon. \tag{D.6}$$

In particular, every contiguous window with $N \leq M$ is admissible.

Proof. For $\Delta \in \mathbb{Z}$, set $\chi_\Delta(g^u) = e^{2\pi \mathrm{i} \Delta u / M}$.

$$\begin{aligned}
g^u = g^v \iff u - v \in M\mathbb{Z} \implies e^{2\pi \mathrm{i} \Delta (u - v)/M} = 1,\\
\chi_\Delta(g^u g^v) = e^{2\pi \mathrm{i} \Delta (u+v)/M} = e^{2\pi \mathrm{i} \Delta u / M} e^{2\pi \mathrm{i} \Delta v / M} = \chi_\Delta(g^u)\chi_\Delta(g^v),\\
\chi_\Delta \equiv 1 \iff \chi_\Delta(g) = 1 \iff e^{2\pi \mathrm{i} \Delta / M} = 1 \iff M \mid \Delta.
\end{aligned}$$

Thus, for $\Delta \not\equiv 0 \pmod M$,

$$\left|\frac{1}{\kappa}\sum_{a\in F} e^{2\pi \mathrm{i}\Delta b_a/M}\right| = \frac{1}{\kappa}\left|\sum_{a\in F}\chi_\Delta(g^{b_a})\right| = \frac{1}{\kappa}\left|\sum_{a\in F}\chi_\Delta(g+a)\right|$$
$$\overset{(D.4)}{\leq} \frac{(\varrho-1)\sqrt{\kappa}}{\kappa} = \frac{\varrho-1}{\sqrt{\kappa}} = \varepsilon,$$
$$v_i^\top v_j = \frac{1}{\kappa}\sum_{a\in F}\big[\cos(i\omega_a)\cos(j\omega_a) + \sin(i\omega_a)\sin(j\omega_a)\big]$$
$$= \frac{1}{\kappa}\sum_{a\in F}\cos((i-j)\omega_a) = \mathrm{Re}\left(\frac{1}{\kappa}\sum_{a\in F} e^{2\pi \mathrm{i}(i-j)b_a/M}\right),$$
$$i-j\not\equiv 0 \pmod M \implies |v_i^\top v_j| \leq \left|\frac{1}{\kappa}\sum_{a\in F} e^{2\pi \mathrm{i}(i-j)b_a/M}\right| \leq \varepsilon,$$
$$j\neq t_a,\quad j-t_a\not\equiv 0 \pmod M \implies q_a^\top(k_{t_a}-k_j) = v_{t_a}^\top v_{t_a} - v_{t_a}^\top v_j = 1 - v_{t_a}^\top v_j \geq 1-\varepsilon.$$

For any nonempty subsets $\mathcal{A}' \subseteq \mathcal{A}$ and $\varnothing \neq D'_a \subseteq J_a \setminus \{t_a\}$,

$$\min_{a\in\mathcal{A}'}\min_{j\in D'_a} q_a^\top(k_{t_a}-k_j) \geq \min_{a\in\mathcal{A}}\min_{j\in J_a\setminus\{t_a\}} q_a^\top(k_{t_a}-k_j) \geq 1-\varepsilon.$$

$$i,j\in\{c,\dots,c+N-1\},\quad i\neq j,\quad N\leq M$$
$$\implies 1\leq |i-j|\leq N-1 < M \implies i-j\not\equiv 0 \pmod M.$$

□

Strict separation and collisions. Independently of the sign of $1-\varepsilon$,

$$\Delta = j-t\not\equiv 0 \pmod M \implies 1-v_t^\top v_j = \frac{2}{\kappa}\sum_{a\in F}\sin^2\left(\frac{\pi\Delta b_a}{M}\right)$$
$$\geq \frac{2}{\kappa}\sin^2\left(\frac{\pi\Delta}{M}\right) \geq \frac{2}{\kappa}\sin^2\left(\frac{\pi}{M}\right) > 0,$$
$$j-t\equiv 0 \pmod M \implies v_j = v_t \implies q^\top(v_t - v_j) = 0 \quad (\forall q).$$

D.3 An elementary verification for quadratic extensions

Proposition D.3 (Exact quadratic character sum magnitude)**.** *For $E=\mathbb{F}_{\kappa^2}$, $g\notin F$, and nontrivial χ on $E^\times$, let $S=\sum_{a\in F}\chi(g+a)$. Then*

$$|S|^2 = \begin{cases}\kappa, & \chi|_{F^\times} \textit{ nontrivial},\\ 1, & \chi|_{F^\times} \textit{ trivial}.\end{cases} \tag{D.7}$$

Proof. Set $\mathcal{D}=\{(a,b)\in F^2 : a\neq b\}$ and $\Phi(a,b)=(g+a)/(g+b)$.

$$\Phi(a,b)=c\in F^\times \implies (1-c)g = cb-a$$
$$\implies (c=1\ \wedge\ a=b) \quad (\perp),$$
$$\Phi(\mathcal{D})\subseteq E^\times\setminus F^\times,$$
$$\Phi(a,b)=\Phi(c,d) \iff (g+a)(g+d)=(g+c)(g+b)$$
$$\iff (a+d-c-b)g + (ad-cb) = 0$$
$$\implies a+d=c+b,\quad ad=cb$$
$$\implies c=a+d-b,\quad 0=ad-b(a+d-b)=(a-b)(d-b)$$
$$\implies d=b,\quad c=a,$$
$$|\mathcal{D}| = \kappa(\kappa-1) = (\kappa^2-1)-(\kappa-1) = |E^\times\setminus F^\times|.$$

Hence $\Phi : \mathcal{D} \longrightarrow E^\times \setminus F^\times$ is bijective. For a finite group H and a multiplicative character ψ,

$$\psi \not\equiv 1 \implies \exists u \in H : \psi(u) \neq 1,$$
$$T_H := \sum_{x\in H} \psi(x) = \sum_{x\in H} \psi(ux) = \psi(u)T_H \implies (1-\psi(u))T_H = 0 \implies T_H = 0,$$
$$\psi \equiv 1 \implies T_H = |H|.$$

Since $|\chi(x)| = 1$, $\overline{\chi(x)} = \chi(x)^{-1} = \chi(x^{-1})$; therefore

$$\begin{aligned}
|S|^2 &= \sum_{a,b\in F} \chi(g+a)\overline{\chi(g+b)} = \sum_{a,b\in F} \chi\left(\frac{g+a}{g+b}\right) \\
&= \sum_{a\in F} \chi(1) + \sum_{(a,b)\in\mathcal{D}} \chi(\Phi(a,b)) = \kappa + \sum_{x\in E^\times\setminus F^\times} \chi(x) \\
&= \kappa + \underbrace{\sum_{x\in E^\times} \chi(x)}_{0} - \sum_{x\in F^\times} \chi(x) = \begin{cases} \kappa, & \chi|_{F^\times} \not\equiv 1, \\ \kappa - (\kappa-1) = 1, & \chi|_{F^\times} \equiv 1, \end{cases} \\
|S| &\leq \sqrt{\kappa}.
\end{aligned}$$

□

D.4 Free low dimensional optimum and asymptotic optimality

Lemma D.4 (Dimension dependent free upper bound)**.** *If $N \geq d+2$, then $\Gamma_{N,d} \leq 1$ under complete competition, even with arbitrary free unit keys and queries.*

Proof. For $d+2$ keys, choose a nonzero affine dependence:

$$\sum_j c_j(k_j, 1) = 0, \quad A = \{j : c_j > 0\}, \quad B = \{j : c_j < 0\}, \quad s = \sum_{j\in A} c_j = -\sum_{j\in B} c_j > 0.$$

Then

$$p = \sum_{j\in A} \frac{c_j}{s} k_j = \sum_{j\in B} \frac{-c_j}{s} k_j \in \operatorname{conv}\{k_j : j \in A\} \cap \operatorname{conv}\{k_j : j \in B\}, \qquad A \cap B = \varnothing.$$

For $p \neq 0$,

$$\|p\|^2 = \sum_{j\in A} \frac{c_j}{s} k_j^\top p \implies \exists t \in A : \; k_t^\top p \geq \|p\|^2,$$
$$p \in \operatorname{conv}\{k_j : j \in B\} \subseteq C_t \implies h_t^2 \leq \|k_t - p\|^2 = \|k_t\|^2 - 2k_t^\top p + \|p\|^2 \leq 1 - \|p\|^2 < 1.$$

For $p = 0$, any $t \in A$ satisfies $0 \in C_t$, so $h_t \leq \|k_t\| \leq 1$. In both cases, Theorem A.2 gives

$$\sup_{\{\|k_j\|\leq 1\}} \min_t h_t = \Gamma_{N,d} \leq 1.$$

□

Theorem D.5 (Finite comparison and deterministic asymptotics)**.** *For $2\kappa \leq d$, $N \leq \kappa^\varrho - 1$, and $\varepsilon = (\varrho-1)/\sqrt{\kappa}$, the embedded Geometric RoPE code has margin at least $1-\varepsilon$. If $N \geq 3$ and $\varepsilon < \cos 1$, it strictly exceeds every code in the standard fixed class. If also $N \geq d+2$ and $\varepsilon < 1$,*

$$0 \leq \Gamma_{N,d} - \gamma_{\mathrm{Geo}} \leq \varepsilon, \qquad 0 \leq 1 - \gamma_{\mathrm{Geo}}/\Gamma_{N,d} \leq \varepsilon. \tag{D.8}$$

Choose $\kappa = 2^{\lfloor \log_2(d/2) \rfloor}$ for $d \geq 4$ and $\varrho = \max(2, \lceil \log_\kappa(N+1) \rceil)$. For sequences satisfying

$$d \to \infty, \qquad d/N \to 0, \qquad \frac{\log(N+1)}{\sqrt{d}\log d} \to 0, \tag{D.9}$$

the additive and relative gaps in (D.8) vanish.

Table 2: **Precomputed frequency certificates.** For each (d, ς), the first row maximizes the certified gap and the second maximizes M, subject to certification and resource constraints; identical choices are merged. Here $\mathbf{gap} = (1 - \varsigma/d)(1 - (\varrho - 1)/\sqrt{\kappa})$. The target zero-frequency fraction $1/4$ is quantized to the nearest prime-power configuration.

d	κ	ϱ	ς	$M = \kappa^\varrho - 1$	$\mathbf{gap} \le \gamma_{\mathrm{Geo}}$	$\mathbf{gap} - (1 - \cos 1)$
No zero frequencies						
256	128	2	0	16383	0.9116	0.4519
256	128	4	0	268435455	0.7348	0.2751
512	256	2	0	65535	0.9375	0.4778
512	256	3	0	16777215	0.8750	0.4153
1024	512	2	0	262143	0.9558	0.4961
1024	512	3	0	134217727	0.9116	0.4519
Target zero-frequency fraction: $1/4$						
1024	383	2	258	146688	0.7098	0.2501
1024	383	3	258	56181886	0.6716	0.2119

Proof. Theorems D.1 and D.2 give

$$\varepsilon < \cos 1 \implies \gamma_{\mathrm{Geo}} \ge 1 - \varepsilon > 1 - \cos 1 \ge \Gamma^{\mathrm{std}}_{N,d}.$$

For $N \ge d + 2$ and $\varepsilon < 1$, Lemma D.4 yields

$$\begin{aligned} 0 < 1 - \varepsilon \le \gamma_{\mathrm{Geo}} \le \Gamma_{N,d} \le 1, \\ 0 \le \Gamma_{N,d} - \gamma_{\mathrm{Geo}} \le 1 - (1 - \varepsilon) = \varepsilon, \\ 0 \le 1 - \frac{\gamma_{\mathrm{Geo}}}{\Gamma_{N,d}} \le 1 - \gamma_{\mathrm{Geo}} \le \varepsilon. \end{aligned}$$

For $\kappa = 2^{\lfloor \log_2(d/2) \rfloor}$ and $\varrho = \max\{2, \lceil \log_\kappa(N + 1) \rceil\}$,

$$\begin{aligned} d/4 < \kappa \le d/2, \qquad \kappa^\varrho \ge N + 1, \qquad \varrho - 1 \le 1 + \frac{\log(N+1)}{\log \kappa}, \\ d \ge 16 \implies \log \kappa \ge \tfrac{1}{2} \log d, \qquad \frac{1}{\sqrt{\kappa}} \le \frac{2}{\sqrt{d}}, \\ 0 \le \varepsilon = \frac{\varrho - 1}{\sqrt{\kappa}} \le \frac{2}{\sqrt{d}} + \frac{4 \log(N+1)}{\sqrt{d} \log d} \xrightarrow[(D.9)]{} 0, \\ d/N \to 0, \quad d \to \infty \implies N \ge d + 2 \quad \text{eventually}. \end{aligned}$$

□

For fixed $a > 1$, $N = \lceil d^a \rceil$ satisfies these conditions and gives error $O_a(d^{-1/2})$. Dimension must grow to maintain a fixed positive margin.

D.5 Ready-to-use frequency allocations

Table 3 lists the complete integer frequencies for the maximum certified period of every recommended dimension layout. For each row, use

$$\kappa = \frac{d_{\mathrm{pos}}}{2}, \quad \varsigma = d - d_{\mathrm{pos}}, \qquad \boxed{\omega_\ell = \frac{2\pi b_\ell}{M} \text{ radians per position}}, \qquad \theta_{t,\ell} = \frac{2\pi((t b_\ell) \bmod M)}{M}. \tag{D.10}$$

Equivalently, $\omega^\circ_\ell = 360 b_\ell / M$ degrees per position.

Table 3: Complete maximum M allocations for all 26 recommended layouts. Each list b contains exactly $d_{\text{pos}}/2$ integers in increasing order; continued rows concatenate the same list. The remaining $d - d_{\text{pos}}$ coordinates have zero frequency.

d	d_{pos}	M	b (increasing)
8	8	15	[1, 2, 4, 8]
16	14	48	[1, 12, 22, 29, 31, 34, 35]
16	16	63	[1, 6, 8, 14, 38, 48, 49, 52]
32	22	120	[1, 30, 38, 55, 56, 65, 69, 71, 76, 99, 118]
32	26	168	[1, 8, 9, 20, 24, 30, 33, 60, 77, 95, 97, 138, 143]
32	32	4095	[1, 210, 448, 601, 697, 865, 1000, 1089, 1262, 1548, 1929, 2159, 3534, 3680, 3766, 3891]
64	50	15624	[1, 1400, 2911, 3953, 4062, 5738, 6422, 6496, 6564, 7652, 7681, 7818, 7843, 8593, 8657, 8697, 8966, 9350, 9762, 13340, 13883, 13942, 14081, 14621, 15555]
64	58	24388	[1, 231, 468, 1416, 2232, 4351, 4565, 4972, 5091, 6382, 6580, 7013, 7548, 7770, 7892, 9640, 9673, 9802, 9955, 10342, 11013, 13949, 14720, 15266, 16442, 18417, 19025, 20693,
64	58	24388	22294]
64	64	1048575	[1, 21021, 50348, 66468, 123520, 126101, 178999, 212012, 219098, 235409, 241370, 304813, 307140, 316858, 345449, 357197, 366603, 394546, 429175, 535448, 578057, 585358, 627172, 691288, 748500, 825689, 873156, 875950,
64	64	1048575	887446, 923796, 952630, 1015891]
128	82	1680	[1, 28, 29, 105, 107, 117, 174, 180, 187, 218, 242, 308, 341, 349, 413, 506, 543, 597, 682, 699, 881, 920, 939, 973, 991, 1034, 1151, 1198,
128	82	1680	1234, 1238, 1249, 1254, 1263, 1284, 1378, 1381, 1404, 1426, 1500, 1600, 1649]
128	98	117648	[1, 911, 3112, 4699, 11447, 11879, 14677, 15641, 17046, 18353, 19615, 20872, 21282, 23202, 23509, 25985, 28724, 30431, 33047, 34171, 34916, 36015, 36417, 39231, 39265, 41392, 48064, 50934,
128	98	117648	52501, 58929, 59167, 64374, 66523, 67408, 74056, 75164, 75508, 77572, 79167, 80833, 83855, 92020, 99737, 102761, 103028, 103185, 114847, 114977, 117282]
128	118	12117360	[1, 37572, 293342, 515094, 543971, 780062, 1143647, 1475108, 1573043, 1662683, 1713437, 1938260, 2107084, 2192728, 2277081, 2319960, 2324172, 2545997, 2593487, 2649971, 2834894, 2845518, 3379748, 3665189, 3827564, 4342278, 4544525, 5029888,
128	118	12117360	5360397, 5723757, 5767298, 5797190, 6010334, 6067949, 6173722, 6189027, 6431974, 6630703, 7120663, 7194245, 7277416, 7337521, 7978550, 8472801, 8724346, 9006946, 9446878, 9595827, 9722656, 10004014, 10055858, 10061660, 10206478, 10345124, 10558614, 10908100,
128	118	12117360	11189620, 11265082, 11795749]
128	128	1073741823	[1, 701065, 7217733, 15849536, 22113942, 34231907, 53859283, 56562413, 69741205, 83872646, 119283668, 188451886, 189534916, 206692829, 227120738, 245477950, 288754667, 308835997, 316383209, 323455154, 330428545, 337626129, 340705101, 356742247, 358814607, 367823492, 376813975, 382781830,
128	128	1073741823	480442558, 500628166, 533915471, 538761218, 542435971, 549623125, 554905983, 556893269, 570004948, 573382983, 635945513, 694866642, 708866614, 718310148, 729724511, 741185116, 756367852, 804992529, 835494886, 900618340, 907637087, 912695575, 921365175, 959609087, 960504620, 967195413, 971376424, 975716690,
128	128	1073741823	991412340, 997824812, 1004600289, 1018722516, 1020077238, 1043389842, 1062526417, 1064430491]
256	162	531440	[1, 345, 18129, 20455, 20545, 31228, 41577, 47111, 64828, 71845, 78161, 79063, 81080, 81897, 82486, 89948, 99115, 103376, 109689, 120813, 121812, 125345, 136033, 138949, 140164, 141394, 141558, 170694,
256	162	531440	180007, 182875, 185253, 185288, 188806, 189016, 194259, 196854, 216235, 228843, 231083, 247428, 249019, 251794, 255177, 259300, 277684, 285851, 285905, 293647, 300006, 302317, 323215, 327350, 328869, 332591, 339057, 345357,
256	162	531440	357163, 359777, 364114, 367486, 375076, 385938, 391157, 395277, 398346, 399088, 412522, 425667, 426130, 435335, 441015, 458080, 459921, 461022, 461072, 476977, 511661, 521989, 522614, 526365, 527192]
256	194	88529280	[1, 115655, 921070, 1123708, 1917022, 2032281, 3388324, 4102125, 4173936, 5273061, 6710939, 6778093, 7477237, 7972247, 9123803, 9318196, 9828144, 11276308, 12675496, 14908021, 15637741, 16534198, 16725467, 16809973, 16842903, 17611245, 18102231, 18435209,
256	194	88529280	21551518, 21656685, 23044993, 25450769, 27870737, 28622352, 28856102, 32026598, 32191486, 32253911, 32501894, 34742830, 34785600, 37682692, 37821399, 38847761, 40947268, 41429128, 42392685, 42941399, 43672089, 44106349, 44274049, 44668554, 45904848, 46708197, 49226880, 50426375,

Table 3 continued

d	d_{pos}	M	b (increasing)
256	194	88529280	50950445, 51727767, 53184153, 54464356, 56301713, 56510152, 59066344, 60823200, 61311978, 62242827, 62344293, 65836683, 66113051, 66397057, 67211470, 68588662, 70111175, 70836503, 71877162, 72286601, 72338830, 74452166, 75183604, 75289534, 76487520, 76687126, 76904284, 77615495,
256	194	88529280	78260077, 78447051, 79801174, 82098789, 82197066, 82248892, 83454709, 83960139, 84026633, 84384137, 84969172, 86469657, 86867356]
256	226	163047360	[1, 4144888, 4623550, 6117618, 12286514, 15225224, 16445456, 16910045, 17206552, 17727269, 19880750, 19992104, 20496069, 21834879, 25489417, 27203965, 27440765, 28533896, 29812966, 30718562, 31195141, 31220722, 31782591, 34893698, 38754373, 40175128, 40761953, 43668445,
256	226	163047360	44982724, 45152703, 47700934, 49804371, 50029269, 50077156, 50404502, 53302754, 59595984, 61926685, 62392767, 65010387, 68551664, 69341492, 70613074, 74080910, 74154597, 79863234, 80108031, 81536449, 81558430, 83875325, 84254820, 84661096, 85703061, 85731750, 87374917, 87578629,
256	226	163047360	87964786, 88127698, 88549119, 90166272, 90264390, 96489310, 97702482, 100923327, 102830580, 104809620, 104815941, 105413493, 105577408, 105934475, 108655350, 108900235, 113358793, 113509988, 115061704, 116460626, 117628030, 117936395, 120614256, 123728417, 123873952, 124759042, 125045523, 125364543,
256	226	163047360	125893070, 126212270, 130656255, 130738552, 131183645, 131285861, 131701590, 132611369, 137241888, 139955816, 141818213, 144708866, 144838980, 145553253, 148679798, 148832979, 149260841, 149696404, 150458997, 150705076, 150718808, 154305317, 155870228, 156969186, 157023778, 158075965, 160738315, 161992952,
256	226	163047360	162577412]
256	256	268435455	[1, 1659325, 4562963, 5103182, 6754155, 8296136, 11776800, 19480442, 22578422, 31895980, 32507226, 32797858, 36471923, 36505151, 37146055, 37225867, 37905233, 41829749, 43249058, 44299071, 46017598, 46025480, 51143891, 56060761, 60696288, 61040235, 63572515, 64668798,
256	256	268435455	70715143, 70993065, 71728046, 78815979, 79474187, 82258489, 82262048, 83482246, 83540253, 83656756, 83672255, 84895311, 85325385, 86956863, 87285770, 96663781, 97381410, 97472674, 97703249, 99803037, 101113688, 101998143, 105212985, 106530030, 107630368, 110858628, 111275047, 111914899,
256	256	268435455	112595404, 114753130, 114827983, 114859559, 117095305, 121404910, 121438700, 125331048, 126010311, 129533586, 136983470, 137831093, 140082691, 140935416, 146780933, 147438653, 147530233, 151990217, 152978837, 153004207, 156654369, 157950789, 160046806, 160441865, 161528940, 162462464, 165839879, 166188338,
256	256	268435455	169088074, 171837038, 173623688, 173770934, 174796882, 176851067, 178194615, 178774268, 186816923, 187584314, 188919289, 189093482, 194311378, 194514173, 196213776, 196485431, 199123334, 201221908, 204241551, 207989058, 208082209, 209849800, 220188883, 220319506, 220390664, 226516222, 226948632, 230093996,
256	256	268435455	230609571, 235706549, 240848116, 245628077, 245847217, 247023617, 247245501, 251815551, 253191442, 254477518, 257174217, 257494573, 260394653, 260799430, 265170464, 266145317]
512	326	705911760	[1, 1345785, 10894065, 17527799, 41691013, 49025991, 52037334, 59122832, 81624845, 94315527, 97186958, 102558435, 109171511, 113469381, 113523662, 120537984, 121126390, 122625286, 126855567, 127550482, 128651718, 129912708, 131186002, 141030504, 143694230, 147293217, 147914484, 148729384,
512	326	705911760	149616151, 160389224, 170375567, 171619164, 177511095, 183536825, 184951938, 193045342, 194377893, 195530209, 196537407, 203005086, 204396116, 204542365, 205144805, 207427577, 209707317, 211016755, 215834007, 217459661, 227378941, 232215591, 233803046, 236020761, 236698110, 236819232, 236878163, 240071234,
512	326	705911760	241073513, 249605175, 253496323, 258445077, 264318552, 265504369, 269023691, 282056048, 285735078, 288439255, 288656681, 292974572, 294600148, 296461832, 296544726, 296951628, 297060539, 311138834, 317037321, 318827627, 319415256, 327381197, 330111550, 335556579, 338582174, 342400138, 352049151, 353079081,
512	326	705911760	377840622, 382465665, 392159867, 397386922, 414127257, 414776649, 422558178, 434834038, 435008842, 440375495, 443041151, 444506768, 446749768, 448051974, 448298387, 451763775, 453700323, 461640385, 478992235, 482720300, 486345839, 488130415, 494361777, 495880632, 498369018, 500921247, 507274488, 507497049,
512	326	705911760	509172326, 517449693, 518471083, 523033539, 524648038, 528690269, 529030458, 538798632, 542798243, 543766439, 552934264, 553664388, 555887775, 558531982, 558755133, 559604153, 564196023, 566413436, 566577017, 568277923, 569024938, 571552384, 572893278, 574503914, 594077319, 605615069, 611100628, 612762001,
512	326	705911760	619943910, 626364174, 626584431, 627229412, 633257536, 633588308, 635083039, 643491760, 656271029, 662332089, 665015888, 668467531, 668503760, 670666522, 673689350, 679000966, 681559554, 683564176, 688819731, 693607460, 694339865, 696641634, 698040719]

Table 3 continued

d	d_{pos}	M	b (increasing)
512	386	7189056	[1, 2580, 3996, 12779, 69995, 133375, 135945, 167235, 170293, 227567, 233635, 273180, 305985, 382496, 393999, 425564, 453216, 576031, 626579, 652465, 657884, 709342, 729730, 749235, 781039, 812794, 836389, 867493,
512	386	7189056	941791, 953101, 989492, 1000386, 1011506, 1123779, 1175941, 1194207, 1208431, 1216109, 1237893, 1280864, 1296616, 1315441, 1388947, 1407209, 1534767, 1562984, 1645372, 1660148, 1680576, 1693748, 1699525, 1713973, 1754895, 1777855, 1825741, 1849474,
512	386	7189056	1911389, 1948531, 1974598, 2067525, 2077579, 2098206, 2104453, 2110950, 2140850, 2196351, 2197585, 2262893, 2273865, 2288386, 2361308, 2409241, 2429862, 2433601, 2487577, 2518195, 2653591, 2735475, 2817471, 2822894, 2844031, 2904152, 2965318, 2987277,
512	386	7189056	2998637, 3127997, 3142511, 3178315, 3199913, 3228313, 3242990, 3268557, 3274344, 3281478, 3323156, 3422632, 3460214, 3502714, 3508499, 3544155, 3559023, 3676567, 3757970, 3762690, 3923627, 3967528, 3986258, 4031798, 4104168, 4117097, 4126013, 4155580,
512	386	7189056	4291274, 4335648, 4508264, 4534143, 4547362, 4547672, 4563055, 4615421, 4617970, 4772305, 4792897, 4870150, 4892917, 4895710, 4896973, 4897218, 4911506, 5011880, 5041044, 5045242, 5103646, 5116233, 5224863, 5241902, 5290644, 5368738, 5377857, 5428568,
512	386	7189056	5467988, 5481030, 5563932, 5578799, 5601807, 5613939, 5636352, 5648401, 5735074, 5744746, 5778711, 5780247, 5840262, 5854299, 5908850, 5990619, 5994720, 6010371, 6031036, 6076695, 6079043, 6100968, 6161747, 6188266, 6212685, 6231875, 6243870, 6251439,
512	386	7189056	6267146, 6308707, 6335361, 6335828, 6414058, 6467205, 6467430, 6503645, 6511670, 6512791, 6590474, 6723423, 6723872, 6749923, 6831967, 6836179, 6864403, 6870565, 6972227, 7028275, 7048925, 7055857, 7061184, 7137739, 7159326]
512	446	11089566	[1, 103797, 109398, 122202, 144239, 152380, 175256, 180684, 199824, 247115, 279113, 296053, 319823, 350320, 373195, 403286, 407239, 425169, 437290, 478854, 565073, 613486, 776314, 777417, 917756, 1265313, 1376301, 1377644,
512	446	11089566	1406098, 1566339, 1588017, 1603088, 1603547, 1668178, 1793318, 1807169, 1814039, 1827512, 1902750, 1962394, 1981116, 2073715, 2237734, 2364061, 2418179, 2516494, 2553052, 2587820, 2593121, 2637363, 2669280, 2680394, 2702357, 2705583, 2713173, 2714585,
512	446	11089566	2753611, 2819883, 2837204, 2852904, 2869033, 2958201, 2960751, 3173560, 3178945, 3234430, 3240175, 3266789, 3338826, 3355337, 3512971, 3625773, 3645613, 3674818, 3680734, 3696745, 3872320, 3899010, 3908517, 4010667, 4023639, 4045282, 4081473, 4119065,
512	446	11089566	4122950, 4152622, 4154933, 4190920, 4287752, 4370860, 4382060, 4408564, 4451743, 4466222, 4520589, 4664121, 4740535, 4844428, 4884593, 4889484, 4912772, 4935891, 4979610, 5014094, 5050040, 5112701, 5192952, 5253798, 5305073, 5433757, 5762831, 5787274,
512	446	11089566	5794291, 5845339, 5913144, 6080598, 6136585, 6162542, 6183293, 6221272, 6257381, 6299792, 6319034, 6324958, 6411741, 6421391, 6520568, 6533181, 6606264, 6661639, 6718764, 6742824, 6748138, 6748528, 6753591, 6948017, 7103470, 7114633, 7129903, 7211655,
512	446	11089566	7262489, 7273986, 7294134, 7327057, 7429108, 7442773, 7459252, 7464031, 7464677, 7470833, 7485569, 7539562, 7703591, 7722575, 7729001, 7753495, 7788852, 7825645, 7856953, 7862739, 7868968, 7945737, 8007777, 8042272, 8138359, 8189604, 8204967, 8316064,
512	446	11089566	8381015, 8409954, 8479655, 8622171, 8743922, 8768436, 8817310, 8935751, 9002843, 9090951, 9213573, 9225121, 9292045, 9293347, 9297250, 9324560, 9328795, 9335961, 9433367, 9433759, 9463837, 9509912, 9543509, 9615779, 9659601, 9672523, 9727897, 9744538,
512	446	11089566	9792681, 9809276, 9812603, 9897537, 9932704, 9994705, 10079220, 10212413, 10254864, 10268965, 10309665, 10396384, 10471054, 10490891, 10494382, 10552183, 10638222, 10651271, 10674870, 10717023, 10720356, 10819435, 10889046, 10909983, 10924486, 10951321, 11025408]
512	512	16777215	[1, 34617, 90740, 320311, 370464, 433697, 571428, 617364, 631231, 723110, 821207, 854770, 902598, 910390, 927915, 930195, 931087, 993037, 1060601, 1095773, 1267257, 1290940, 1368700, 1384902, 1479907, 1547113, 1555424, 1560792,
512	512	16777215	1563376, 2022388, 2060587, 2230804, 2295749, 2505695, 2520944, 2528077, 2727992, 2750692, 2936281, 2994662, 3003324, 3042176, 3057027, 3111336, 3154170, 3184373, 3195887, 3255473, 3273055, 3273773, 3313112, 3362288, 3487898, 3500074, 3625208, 3643176,

Table 3 continued

d	d_{pos}	M	b (increasing)
512	512	16777215	3788931, 3809712, 3822774, 3860609, 3934655, 3989095, 4001518, 4028643, 4111244, 4184753, 4524563, 4545695, 4576238, 4608304, 4691576, 4707835, 4728659, 4763490, 4816680, 4864794, 4982704, 4984429, 5054591, 5090049, 5096489, 5114717, 5135088, 5154748,
512	512	16777215	5213048, 5314687, 5361398, 5394866, 5643111, 5695735, 5762518, 5827368, 5923228, 5926804, 5940411, 5963999, 5981770, 6324735, 6346282, 6373132, 6378778, 6528682, 6558015, 6674588, 6823235, 6837668, 6843202, 6910382, 6912145, 7351011, 7437907, 7459352,
512	512	16777215	7499291, 7550775, 7563074, 7832770, 7871593, 7883131, 7919916, 7948492, 7988989, 8060444, 8166599, 8196118, 8230947, 8245360, 8360907, 8443549, 8461006, 8494433, 8583263, 8605384, 8651758, 8680752, 8738043, 8755585, 8847605, 8989897, 8990685, 9017580,
512	512	16777215	9045964, 9088198, 9162801, 9239423, 9362039, 9404603, 9501642, 9532862, 9558906, 9598514, 9768592, 9796206, 9919516, 9942537, 10039656, 10059653, 10393217, 10457501, 10535065, 10617766, 10632390, 10759161, 10950112, 11008558, 11036847, 11050710, 11148541, 11251878,
512	512	16777215	11252167, 11296496, 11309788, 11338607, 11515778, 11719740, 11743016, 11784701, 11862698, 11872644, 11889154, 11891899, 11943064, 11958997, 12041812, 12052950, 12208607, 12236481, 12292901, 12295449, 12348971, 12564082, 12614076, 12635424, 12754402, 12827003, 13074823, 13079078,
512	512	16777215	13102296, 13140432, 13199755, 13217450, 13233971, 13243607, 13369962, 13375770, 13476302, 13524619, 13533059, 13714495, 13751213, 13840619, 13851926, 13986209, 14002370, 14015801, 14029510, 14037556, 14041647, 14071577, 14135060, 14140741, 14162422, 14308711, 14495471, 14562386,
512	512	16777215	14633736, 14659036, 14661447, 14739687, 14826072, 14827148, 14842576, 15029994, 15191797, 15306372, 15647116, 15650503, 15703722, 15707849, 15786264, 15793572, 15862238, 15890143, 15942316, 15961968, 16062565, 16153906, 16187897, 16307732, 16365374, 16399100, 16463053, 16543838,
512	512	16777215	16598229, 16665777, 16691422, 16705041]
1024	512	65535	[1, 60, 256, 586, 1615, 1723, 2481, 2607, 3052, 3362, 3803, 3871, 4394, 4398, 4422, 4526, 4584, 4815, 5010, 5271, 5294, 5464, 5878, 6009, 6275, 6355, 6386, 6455,
1024	512	65535	6553, 6718, 7951, 8106, 8622, 8624, 8636, 8688, 8717, 8930, 9072, 9157, 9364, 9618, 9784, 10478, 10736, 10737, 10743, 10769, 10890, 10961, 11107, 11234, 11317, 11664, 11793, 11870,
1024	512	65535	12042, 12257, 12360, 12446, 12605, 12648, 12749, 13231, 13612, 13949, 14105, 14374, 14695, 15048, 15360, 15665, 15898, 16535, 16540, 16837, 17246, 17371, 17651, 17753, 17870, 17937, 18480, 18946,
1024	512	65535	19297, 19651, 19911, 20220, 20230, 20824, 21607, 21642, 21892, 22452, 22549, 22609, 22656, 22853, 22890, 23024, 23205, 23657, 24110, 24131, 24907, 25042, 25387, 25744, 25763, 25785, 26425, 26452,
1024	512	65535	26673, 26972, 26989, 27225, 27339, 27590, 27610, 27635, 28009, 28707, 28798, 29329, 29695, 30155, 30364, 30434, 30934, 30999, 31423, 32054, 32248, 32368, 32462, 32856, 32930, 33198, 33445, 33560,
1024	512	65535	33771, 33877, 34464, 35370, 35412, 36909, 36964, 37234, 37395, 37413, 37563, 37924, 38638, 38676, 38720, 39193, 40000, 40054, 40433, 40496, 41094, 41128, 41600, 41828, 42197, 42330, 42370, 42420,
1024	512	65535	42469, 43168, 43551, 43771, 44561, 44564, 44577, 44673, 44746, 44851, 45089, 45321, 45495, 45567, 45808, 46167, 46245, 46540, 47025, 47460, 47611, 47878, 48018, 48069, 48161, 48841, 49018, 49148,
1024	512	65535	49996, 50467, 50497, 50619, 50795, 51021, 51258, 51646, 51886, 52074, 52085, 52405, 52529, 52687, 52805, 52862, 53010, 53197, 53546, 53857, 54040, 54270, 54692, 54904, 55915, 56078, 56131, 57647,
1024	512	65535	57890, 57899, 57974, 58789, 59409, 60291, 60427, 60968, 61078, 61473, 61481, 61529, 61618, 61737, 61853, 61940, 61976, 62276, 62315, 62705, 62998, 63227, 63273, 63613, 64426, 64441, 64502, 64590,
1024	512	65535	64703, 65235, 65395, 65457]
1024	634	31855012	[1, 188829, 341000, 537492, 619952, 802644, 830007, 955166, 1056994, 1201804, 1461429, 1483879, 1566168, 1614110, 1802854, 1816821, 2123591, 2238611, 2263898, 2307402, 2445310, 2491289, 2681259, 3168932, 3378159, 3387674, 3392966, 3525690,
1024	634	31855012	3700878, 3783072, 3801417, 3995248, 4041635, 4254018, 4463637, 4578274, 4866326, 4899591, 4947719, 5206728, 5251856, 5256275, 5283314, 5413760, 5498872, 5591917, 5631969, 5691697, 5699152, 5717045, 5822243, 6049064, 6081565, 6097134, 6138820, 6238397,

Table 3 continued

d	d_{pos}	M	b (increasing)
1024	634	31855012	6402898, 6408329, 6669099, 6673432, 6723968, 6906700, 6941695, 7066320, 7087564, 7194169, 7263607, 7323021, 7518007, 7519108, 7530101, 7671690, 7722441, 7760597, 7796735, 7895112, 8274050, 8294078, 8452154, 8669793, 8850303, 8935672, 8979263, 9017861,
1024	634	31855012	9024388, 9101823, 9135562, 9135684, 9393754, 9813550, 9852251, 10061614, 10125684, 10149777, 10181806, 10202595, 10367881, 10674872, 10702011, 10800837, 10845768, 11177355, 11197531, 11262514, 11296190, 11466743, 11647722, 11679003, 11758606, 12025401, 12094988, 12309299,
1024	634	31855012	12473461, 12658433, 12824170, 12870400, 13012655, 13054791, 13125028, 13275110, 13433602, 13607450, 13633844, 13734309, 13936972, 14018209, 14316651, 14386396, 14425176, 14452393, 14642943, 14670153, 14909891, 14983423, 14996881, 15097742, 15213873, 15267962, 15284676, 15298657,
1024	634	31855012	15385221, 15550533, 15668158, 15736308, 15754526, 15766371, 16341594, 16428490, 16639319, 16694115, 16789066, 16791741, 16869334, 16872799, 16927318, 17025494, 17096680, 17119247, 17129507, 17196445, 17298870, 17317058, 17364228, 17437469, 17696460, 17996999, 18011419, 18073455,
1024	634	31855012	18111602, 18221509, 18268320, 18363436, 18397994, 18465038, 18802891, 18898814, 19165491, 19295807, 19459499, 19461702, 19643862, 19697977, 19759905, 20041822, 20094712, 20229221, 20783625, 20869694, 20926329, 21004928, 21046180, 21068655, 21176061, 21285830, 21296012, 21698138,
1024	634	31855012	21712834, 21734263, 21850617, 21856485, 21921998, 21927898, 22001328, 22118289, 22245086, 22262272, 22272664, 22309270, 22602784, 22611251, 22686306, 22689036, 22736050, 22871373, 22960253, 23014738, 23149478, 23184754, 23200390, 23318652, 23418464, 23531969, 23557609, 23594313,
1024	634	31855012	23706833, 23946445, 23949794, 24069960, 24104694, 24257972, 24269651, 24279307, 24313487, 24420316, 24471457, 24484426, 24694502, 24837635, 24883796, 24894273, 25040788, 25203381, 25220295, 25279900, 25611722, 25691187, 25704838, 25726366, 25757187, 25879647, 25940428, 26157334,
1024	634	31855012	26280662, 26358125, 26370434, 26371896, 26433199, 26438044, 26442629, 26553141, 26560305, 26640492, 26762328, 26781893, 27039729, 27127985, 27150317, 27201970, 27378754, 27673696, 27678717, 27721990, 27724438, 27895858, 27896259, 27980438, 27999658, 28260923, 28333315, 28367850,
1024	634	31855012	28572120, 28573735, 28582693, 28681783, 28761699, 28773297, 28864765, 28892457, 28963812, 29090183, 29173356, 29196105, 29226770, 29228653, 29295531, 29403736, 29461415, 29625965, 29629107, 29857669, 30035437, 30043809, 30068873, 30426372, 30549102, 30580462, 30665800, 30896732,
1024	634	31855012	31058188, 31137017, 31157611, 31187389, 31239862, 31276186, 31298851, 31513776, 31581672]
1024	766	56181886	[1, 26224, 111316, 133072, 518196, 670536, 836487, 882935, 949968, 1070858, 1115281, 1140302, 1163102, 1459944, 1847606, 1985910, 2041491, 2434033, 2610349, 2684373, 3071104, 3120520, 3161390, 3190899, 3476629, 3641276, 4278135, 4285921,
1024	766	56181886	4351401, 4377716, 4860413, 4898565, 4922105, 5023074, 5153958, 5298839, 5425781, 5716209, 6086430, 6157893, 6250899, 6268841, 6470374, 6682621, 7393511, 7630567, 7848919, 7875438, 7889395, 7925952, 7986590, 9106406, 9229338, 9438177, 9453668, 9681039,
1024	766	56181886	9906536, 9985224, 10427318, 10459811, 10562002, 10740034, 10934250, 10994887, 11171556, 11198261, 11207607, 11333451, 11385625, 11479041, 11656029, 11701853, 11776815, 11940941, 11991088, 12039841, 12182852, 12256864, 12264038, 12548200, 12583539, 12668338, 12719129, 12806917,
1024	766	56181886	12896685, 12931596, 13043341, 13151366, 13266679, 13299584, 13321038, 13436657, 13475212, 13609791, 13671450, 14026853, 14027637, 14125356, 14151456, 14254970, 14269929, 14561168, 14624457, 14733636, 14895976, 14901917, 15104204, 15153936, 15344822, 15522437, 15803902, 15969663,
1024	766	56181886	16246104, 16633199, 16757347, 17017382, 17060131, 17386306, 17559070, 17615268, 17645699, 17742226, 17831123, 17910139, 18240405, 18373301, 18475709, 18492537, 18551587, 18659947, 19013049, 19169533, 19181469, 19238251, 19871836, 20011617, 20027438, 20032116, 20066375, 20398904,
1024	766	56181886	20668165, 20699819, 20766044, 20845895, 21109739, 21284608, 21311680, 21453926, 21527520, 21642835, 21831365, 22002151, 22079611, 22106549, 22127470, 22148088, 22150252, 22209352, 22361925, 22700796, 22995144, 23059063, 23568323, 23586323, 23793162, 23820322, 23982345, 24562311,
1024	766	56181886	24642034, 24851461, 25007422, 25015960, 25041555, 25246579, 25316242, 25409214, 25425552, 25551340, 26338335, 26598349, 26625004, 26706911, 26728932, 26857233, 26923641, 26988072, 27072676, 27228645, 27374268, 27661398, 27808730, 27829350, 27879611, 28062093, 28090946, 28130559,

Table 3 continued

d	d_{pos}	M	b (increasing)
1024	766	56181886	28158850, 28194258, 28569953, 28752870, 28813440, 29037087, 29076629, 29178090, 29281340, 29353941, 29383433, 29444885, 29457124, 29616377, 29616849, 29631356, 29694181, 29865241, 30002923, 30141665, 30291959, 30441663, 30451699, 30725704, 30835398, 31076036, 31143651, 31145571,
1024	766	56181886	31208714, 31787511, 31803300, 32204062, 32328668, 32333761, 32483708, 32830855, 32934351, 33002765, 33308343, 33486864, 33487295, 33511616, 33543899, 33739550, 33755421, 33864499, 33958008, 34170557, 34173282, 34223373, 34646540, 34823472, 35044388, 35190694, 35251654, 35320483,
1024	766	56181886	35330405, 35627054, 35905791, 35953736, 36015120, 36061942, 36080452, 36206732, 36268168, 36368231, 36485308, 36657788, 37618901, 37633110, 37647082, 38123333, 38195821, 38423075, 38696177, 39022183, 39568123, 39568232, 39683964, 39695590, 39712012, 39780207, 39896491, 40338540,
1024	766	56181886	40396091, 40521418, 40563849, 40565693, 40682676, 40780369, 40833907, 41007886, 41033040, 41137247, 41140962, 41304761, 41364402, 41367380, 41387298, 41499107, 41662952, 42174852, 42308393, 42710759, 42799200, 42830334, 42934727, 43103059, 43245103, 43733878, 44348060, 44460171,
1024	766	56181886	44470022, 44596559, 44615156, 44805160, 44825998, 45004956, 45143306, 45481946, 45512921, 45634795, 45694310, 45957045, 46380808, 46516956, 46966141, 47372879, 47449702, 47537932, 47706419, 47727455, 47780766, 47852833, 47891830, 47915835, 47931219, 48071200, 48092610, 48256235,
1024	766	56181886	48583493, 48834619, 48835851, 48841380, 49167250, 49174049, 49222601, 49481827, 49926826, 50158418, 50351011, 50442251, 50556070, 50591065, 51122953, 51257954, 51456341, 51685382, 51731740, 51808630, 51817761, 51950517, 52009917, 52033050, 52047799, 52283512, 52287899, 52482837,
1024	766	56181886	52872124, 53606322, 53664865, 53806520, 53852761, 53969818, 54166152, 54185075, 54225091, 54511721, 54610378, 54627963, 54658453, 54798451, 55122220, 55365260, 55784448, 55936361, 55936824]
1024	898	90518848	[1, 89242, 153237, 242881, 598075, 635697, 679615, 684726, 738014, 769256, 1023393, 1169800, 1200037, 1256236, 1331712, 1401189, 1680532, 1719825, 2155069, 2283016, 2406021, 2686576, 2747525, 2752515, 2763190, 3395614, 3703320, 3809095,
1024	898	90518848	3936609, 3975771, 3979203, 3993258, 4511133, 4983025, 5564854, 5739846, 5921700, 6254838, 6461930, 6584902, 6737641, 6764183, 6844752, 6909698, 6970264, 7065971, 7121237, 7429191, 7603483, 7818608, 7838190, 8313686, 8446593, 8513433, 8544415, 8852420,
1024	898	90518848	8927039, 9090592, 9195456, 9375470, 9399934, 9667232, 9797447, 9835060, 9857859, 9865850, 10096626, 10141955, 10151497, 10426444, 10451141, 10793086, 11812211, 12049051, 12082125, 12211005, 12408744, 12899064, 12971046, 13010629, 13041279, 13240160, 13415352, 13602332,
1024	898	90518848	14260753, 14275119, 14429414, 14449166, 14672081, 14732713, 14777945, 15030221, 15050436, 15053186, 15440498, 15701828, 15945927, 16363657, 16429354, 16490602, 16492926, 16705810, 16777488, 17269374, 17270206, 17557822, 17825547, 18178335, 18401456, 18864459, 19100532, 19252814,
1024	898	90518848	19457981, 19499331, 19635285, 19643129, 19670138, 19768418, 19958768, 20292580, 20317849, 20386871, 21157889, 21371514, 21744193, 22196469, 22720175, 22999951, 23431713, 23440165, 23502457, 23509740, 24223798, 24437043, 24720807, 24759716, 24987343, 24989172, 24990447, 25112338,
1024	898	90518848	25565975, 25747513, 25753359, 25873352, 26104422, 26155625, 27028432, 27068714, 27143580, 27333401, 27694776, 28004786, 28033266, 28546009, 28727165, 28775889, 29130282, 29526257, 29952499, 29971384, 30363119, 30437472, 30681600, 30717335, 30719326, 30834143, 30863070, 31488834,
1024	898	90518848	31836354, 32199607, 32207916, 32252731, 32288183, 32390057, 32485766, 32527315, 32945207, 32958256, 33057524, 33121355, 33187352, 33534309, 33627410, 33750119, 33844281, 34304403, 34562200, 34712260, 34828775, 34922893, 35810089, 35866823, 36088739, 36167270, 36189758, 36299915,
1024	898	90518848	36317016, 36344186, 36464905, 36642996, 37318661, 37808025, 38322276, 38486706, 38510523, 38586445, 38610871, 38678017, 38749312, 39591863, 39604735, 39626527, 39854752, 40047537, 40317846, 40420167, 40501755, 40556793, 41004818, 41080626, 41197888, 41276059, 41618182, 42055504,
1024	898	90518848	42125862, 42264188, 42351557, 42416070, 42521071, 42841835, 43219055, 43305541, 43534057, 44382642, 45014834, 45259427, 45421063, 45425283, 45822981, 46389887, 46537327, 46582046, 46692448, 46745323, 46760570, 47007910, 47036986, 47094683, 47634296, 47760992, 47809038, 47856474,
1024	898	90518848	47941957, 48271079, 48361229, 48521589, 48542372, 48867444, 48979801, 49253037, 49380163, 49381128, 49950890, 50257258, 50399995, 50481514, 50526737, 51258529, 51709424, 51791045, 52049418, 52520414, 52710110, 52731267, 53164068, 53290415, 53431310, 53536524, 53544818, 53758220,

Table 3 continued

d	d_{pos}	M	b (increasing)
1024	898	90518848	53797777, 53875674, 54042044, 54096214, 54098024, 54167767, 54738522, 55051416, 55104181, 55432271, 55559490, 55576529, 55646274, 56119914, 56141068, 56590883, 56599717, 57287595, 57308631, 57624771, 57727567, 57911180, 58119658, 58156611, 59190517, 59472619, 59647501, 59809452,
1024	898	90518848	60338752, 61188716, 61226973, 61334751, 61479609, 62577223, 63163469, 63184578, 63640511, 63978888, 64077729, 64093106, 64154816, 65194515, 65228244, 65255860, 65318886, 65404788, 65768660, 66238118, 66466646, 66627753, 66660786, 66955047, 67019104, 67160770, 67420390, 67976961,
1024	898	90518848	68172124, 68337216, 68395428, 68982667, 69595265, 69902455, 70302647, 70385114, 70767137, 70796850, 70855542, 70962660, 71387231, 71565621, 71684896, 71792908, 72092795, 72157538, 72193358, 72353222, 72646733, 73750148, 73900420, 73986260, 73996538, 74048773, 74162103, 74643116,
1024	898	90518848	74657042, 74764965, 74944827, 75045195, 75312227, 75677260, 76044278, 76595903, 76698399, 76736243, 76762419, 77160450, 77447048, 77858254, 78083811, 78180778, 78289469, 78601577, 78905955, 79268446, 79610749, 79700245, 80094136, 80328392, 80578815, 81018031, 81074176, 81292892,
1024	898	90518848	81410427, 81471435, 81665089, 82209687, 82246012, 82254782, 82274305, 82325062, 82717032, 82814414, 82994972, 83050210, 83660883, 83791582, 83854634, 83892748, 84046241, 84110957, 85071448, 85378154, 85387388, 85422888, 85570258, 85656907, 85710386, 85753427, 86019714, 86180251,
1024	898	90518848	86312068, 86367504, 86369881, 86393894, 86496924, 86594644, 86710607, 86723040, 86832928, 86868195, 87118949, 87292167, 87346243, 87427249, 87812534, 87909176, 89072407, 89145176, 89321426, 89448258, 89508667, 89596459, 89647008, 90008234, 90034838, 90183676, 90219203, 90248638,
1024	898	90518848	90366723]
1024	1024	134217727	[1, 363087, 507814, 1044293, 1275913, 1332205, 1398014, 1575312, 1745031, 1759546, 1767100, 2293130, 2297951, 2482507, 2941026, 2993305, 4248017, 4591672, 4925442, 5429951, 5835294, 5995082, 6053104, 6282107, 6307623, 7029069, 7146827, 7639008,
1024	1024	134217727	7902676, 8001818, 8214051, 8434732, 8441408, 8665580, 9195596, 9396640, 9399444, 10457064, 10592846, 11057871, 11432581, 11459843, 11701128, 11753953, 12151343, 12254575, 12301001, 12445005, 12471278, 12526995, 12665488, 13402584, 13495377, 13968763, 14927103, 15506729,
1024	1024	134217727	15806114, 16817586, 17238277, 18200690, 18279081, 18999915, 19142847, 19507707, 20066968, 20138749, 20405574, 20741686, 20769550, 20888922, 20932890, 21082124, 21207867, 21498504, 21506494, 21759053, 21831921, 22121231, 23092661, 23216631, 23262478, 23374173, 23567528, 23675345,
1024	1024	134217727	24014342, 24506593, 24560839, 24715898, 24782436, 26178364, 26311259, 26353765, 26772964, 27140562, 29592254, 29931188, 29961019, 30317927, 30328653, 30725773, 31036656, 31400054, 31505877, 32361263, 33030701, 33069614, 33147052, 33177526, 33369458, 33631454, 33903392, 33971584,
1024	1024	134217727	34187317, 34200086, 34484899, 35955629, 36067232, 36069048, 36414709, 36878875, 37588108, 37744678, 37824724, 37903528, 38023320, 38224715, 38474206, 38921485, 39103275, 39121775, 39634464, 39746981, 40254220, 40301241, 40370442, 40641420, 40971810, 41265292, 41414535, 41423457,
1024	1024	134217727	41677633, 42002875, 42026319, 42452859, 42653769, 42677955, 43135429, 43184416, 43230434, 43683020, 43742474, 44133493, 44367762, 44445041, 44565138, 44912137, 45731885, 46306063, 46367729, 46431619, 46584315, 47159760, 47284342, 47408574, 47474143, 47524315, 47857479, 47863991,
1024	1024	134217727	48721286, 48745539, 49043188, 49472578, 50237074, 50738701, 50800503, 50919133, 51249878, 51414942, 51633484, 51928242, 52160896, 52232990, 52308545, 52359147, 52575175, 52842954, 53135350, 53364758, 53365771, 54019857, 54099082, 54308591, 54445705, 55725472, 55859062, 55940767,
1024	1024	134217727	56205363, 57153537, 57361981, 57591206, 57591906, 57975439, 58063712, 58090338, 58367585, 58706664, 58733494, 59194982, 59265855, 59848933, 60175028, 60529871, 60555210, 60602109, 60747755, 61038930, 61112536, 61472572, 62199791, 62287988, 63147236, 63173691, 63516660, 63550423,
1024	1024	134217727	63686940, 63778558, 63796963, 63912979, 64201554, 64514926, 64583047, 64811200, 64882376, 64911250, 65071056, 65136794, 65219704, 65234380, 66053733, 66267991, 66377497, 66650293, 66742367, 67231062, 67241562, 67538537, 68003141, 68611726, 69294192, 69674556, 69861638, 70158783,
1024	1024	134217727	70194403, 70225281, 71134880, 71777438, 72016784, 72348966, 72360153, 72807958, 72849293, 72851195, 72862724, 72918023, 73002793, 73191657, 73203337, 73945104, 74005835, 74229690, 74479692, 74497078, 74586310, 74723590, 75278768, 75397574, 75752277, 75955367, 76289928, 76385780,

Table 3 continued

d	d_{pos}	M	b (increasing)
1024	1024	134217727	76398094, 77622657, 77943119, 78000466, 78078356, 78256308, 78356342, 78542975, 78559293, 78562561, 78849799, 78959733, 79210960, 79451908, 79882202, 79978789, 80299246, 80586453, 80832958, 81437781, 81594764, 81604843, 81670248, 81680134, 82755149, 82795664, 82863597, 82894880,
1024	1024	134217727	83041157, 83291869, 83729701, 83793810, 84438095, 85324831, 85581626, 85782939, 86047544, 86305603, 86378526, 86487136, 86569705, 86730862, 86792622, 87441211, 87494415, 87712743, 88320253, 88359185, 88377466, 88607168, 89246202, 89361311, 89596742, 89754614, 89832027, 89944946,
1024	1024	134217727	90438768, 91179882, 91328137, 91481872, 91875625, 92027524, 92085876, 92561668, 92715927, 93115449, 93248981, 93656803, 93973102, 94219589, 94625078, 94646245, 94664630, 94834325, 94893546, 95369764, 95836617, 96123606, 96790680, 97417202, 98836796, 98844515, 99606128, 99753342,
1024	1024	134217727	99954147, 100557889, 100613992, 100894034, 101324192, 101373384, 101584127, 102267691, 102709907, 102975940, 103016038, 103412219, 103532173, 104000506, 104340316, 104352547, 104496243, 104890145, 104982949, 105009100, 105306150, 105324096, 105522420, 105980764, 106020985, 106158192, 106449677, 106796444,
1024	1024	134217727	107116880, 107282096, 107497738, 107679239, 107737919, 107885964, 108104087, 108248262, 108663823, 108771976, 108946064, 109027527, 109315918, 109335171, 109521312, 109723753, 109744243, 110084870, 110095811, 110133923, 110423988, 110842907, 111032850, 111837138, 112470851, 112630897, 112635980, 112727431,
1024	1024	134217727	113887496, 113910622, 114101717, 114378302, 114858153, 114964644, 115451683, 115457821, 115730433, 115840503, 116084357, 116116798, 116258094, 116783042, 116844941, 116953530, 117180110, 117519814, 117821792, 117969115, 118204077, 118498472, 119244002, 119622426, 120022034, 120276774, 120474821, 120589118,
1024	1024	134217727	120668044, 120668752, 120848786, 121344564, 121807737, 122077246, 122387298, 122387609, 122640872, 122724016, 122753416, 123086336, 123279980, 123296271, 123297046, 123755435, 123823169, 123890966, 124097888, 124238525, 124289848, 124726208, 125027278, 125453636, 126148445, 126170812, 126197123, 126430271,
1024	1024	134217727	127269590, 127434167, 127704768, 128361402, 128611130, 129042895, 129175199, 129208601, 129379624, 129683533, 129730442, 129746126, 130128716, 130469260, 130506982, 130570919, 130670042, 130788739, 130832406, 130846189, 131272739, 131352146, 131487437, 131875494, 132286762, 132699223, 132811326, 133251006,
1024	1024	134217727	133290958, 133454352, 133489180, 133531167, 133558353, 133768320, 133839159, 134021617]

E Experimental protocol and measurement definitions

E.1 Experiments compute resources

Experiments were conducted on a server running Ubuntu 22.04.4 LTS. The system is equipped with an Intel Xeon Gold 6133 processor featuring 80 logical cores at 3.0GHz, and 352 GB of RAM. The machine is configured with 8 NVIDIA GeForce RTX 4080 GPUs with 16GB of video memory each.

E.2 Data generation and archived training configurations

Each sequence has one marker $m \in \{1, \ldots, K\}$ with offset $d_m = 2 + 4(m-1)$ and a legal position $i \in \{d_m, \ldots, L-1\}$. The label at that marker is x_{i-d_m}, other positions have ignored labels. Ordinary tokens are independently sampled, with range $\{100, \ldots, 499\}$. Marker sampling weights are all one. The training and test sets are partitioned according to the congruence relation modulo 8 between the target token ID and the marker's absolute position. For the training set, the target token appearing at the offset position before the marker has an ID that is non congruent with the marker's absolute position modulo 8, whereas the test set comprises the congruent cases. This ensures that, although every token may appear as a target during training, the target conditions of the training and test sets are strictly disjoint. The main twenty marker setting has $K = 20, L = 100$ and legal target windows $0 \le t \le 99 - d_m$. The results presented in the main text are obtained with random seed 1. We conducted experiments across three distinct seeds (1,2,3) in total; results for the remaining two seeds are reported in later sections. Unless otherwise noted, all figures in the paper are generated using the checkpoint at epoch 209, corresponding to the end of training. Table 4 reports detailed configuration.

Table 4: Configuration values.

Item	Value
Sequence length, model width, head width	100, 64, 32
Attention heads and layers	2 heads; 3 layers, except Kimi hybrid: 4
Feed-forward width; activation	64; SiLU
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $eps = 1e-8$, and weight decay=0.01)
Epochs	210 (epochs 0–209)
Initialization rate, gradient clipping	0.5, configured norm threshold 1.0
Learning rate schedule	Initial 10^{-5}; 10-epoch gradual warmup, multiplier 25; cosine schedule $T_{\max} = 200$, floor 10^{-5}
Batch size	1,024
Train / test datasize	2,000,000 / 250,000
RoPE base; pairing	10,000; interleaved pairs
T5-style bias configuration	32 buckets; maximum distance 128
KERPLE configuration	Log form parameters initialized to 1; epsilon 10^{-6}

E.3 Position means, residual content, and legal query supports

At a fixed stage, let $K_{b,j} \in \mathbb{R}^{d_h}$ denote a key under background probe b and physical position j. Here $B_p \in \mathbb{Z}_{>0}$ is the number of independently sampled background token sequences (content probes) used at each fixed position and marker condition, with $b \in \{1, \dots, B_p\}$. The same indexed backgrounds are paired across positions and marker conditions; B_p is a probe count, distinct from the number of distractors B in Proposition A.1 and from the RoPE base. For these equal-size paired probes, define

$$\bar{K}_j = \frac{1}{B_p} \sum_{b=1}^{B_p} K_{b,j}, \quad K_{b,j} = \bar{K}_j + \epsilon_{b,j}, \quad F_K = J_L \bar{K}, \quad J_L = I - \mathbf{1}\mathbf{1}^\top / L. \tag{E.1}$$

For marker m, the query $\bar{Q}_m(t)$ is observed at physical position $i = t + d_m$ but indexed by its target t. Its support is $\mathcal{T}_m = \{0, \dots, L - d_m - 1\}$, center over that support only. Report position means and single-sample residuals separately.

E.4 Rank, effective dimension, and physical-frequency projection

For a nonzero centered matrix F with singular values σ_j, define

$$p_j = \frac{\sigma_j^2}{\sum_r \sigma_r^2}, \quad R_2 = p_1 + p_2, \quad \zeta = \frac{\sigma_2}{\sigma_1}, \quad r_{\text{eff}} = \frac{(\sum_j \sigma_j^2)^2}{\sum_j \sigma_j^4}. \tag{E.2}$$

These quantities are undefined for zero total energy. The Eckart–Young characterization gives $\min_{\text{rank } H \le 2} \|F - H\|_F^2 = (1 - R_2)\|F\|_F^2$. A high R_2 can still describe a line, so ζ is necessary.

On actual observed coordinates $t \in \mathcal{T}$, build the two columns $C_r(t,:) = (\cos(2\pi r t/L), \sin(2\pi r t/L))$. To fit a centered trajectory with an intercept removed, use $B_r = J_{|\mathcal{T}|} C_r$ and

$$\Pi_r = B_r B_r^\dagger, \qquad e_r(F) = \frac{\|\Pi_r F\|_F^2}{\|F\|_F^2}. \tag{E.3}$$

E.5 Empirical margins and their aggregation

For each background sample and valid marker position, unit-normalize each query/key before computing

$$\widehat{\gamma}_{b,m,i} = \widehat{q}_{b,m,i}^\top \widehat{k}_{b,t} - \max_{j \in J_{m,i} \setminus \{t\}} \widehat{q}_{b,m,i}^\top \widehat{k}_{b,j}. \tag{E.4}$$

The maximum is taken within each sample before averaging over backgrounds. The supplied normalized-margin configurations use epsilon 10^{-12}, 64 samples per position, the last layer, and distractors $0, \ldots, i-1$ excluding the target. Backgrounds are paired across markers and positions.

F ADDITIONAL RESULTS

This section tests the geometric account in Section 3 through robustness comparisons, interventions, constrained reconstruction, and frequency diagnostics, before comparing standard and Geometric RoPE.

Unless explicitly stated otherwise, the XoM experiments use the protocol in Section E, the configuration in Table 4, and the final checkpoint at epoch 209. Layer and head indices are zero based.

F.1 MAIN EXPERIMENT RESULTS ACROSS MULTIPLE SETTINGS

F.1.1 RESULTS ACROSS MULTIPLE SEEDS

To establish that the phenomena reported in the main text are robust and general rather than anecdotal, we replicated main text experiments across three independent random seeds. Fig.7, Fig.8, and Fig.9 show the QK score profiles for seeds 1–3. Across the five encodings, the dominant peak generally moves with the prescribed offset $d_m = 2 + 4(m-1)$. As is evident, the phenomena described in the main text are consistently observed across every random seed, every model, and all markers. These are QK diagnostics; for architectures with additive attention biases, the QK component alone does not determine the final attention distribution. Nevertheless, the pronounced similarity of their QK patterns reveals that this is a pivotal mechanism underlying task completion.

Fig.10 and Fig.11 compare the keys' trajectories with the markers' after indexing each query by its target $t = i - d_m$. Most query curves track the key curve on their legal supports. The recurrence of this alignment in both retrieval heads is a strong evidence for a shared positional component.

Fig.12 and Fig.13 provide complementary joint two dimensional projections. They show approximately circular key paths and marker-query arcs. This is consistent with the phenomena presented in the main text, indicating that the key and query likely undergo uniform rotation in two dimensional space.

F.1.2 RESULTS ACROSS MULTIPLE MODELS

In addition to the five non RoPE models shown in the main text, we extend the comparison to fixed sinusoidal embeddings (SinPos) (Vaswani et al., 2017), PaTH attention (Yang et al., 2026), and a small hybrid using Kimi Delta Attention (KDA) (Team et al., 2025). Table 5 specifies the architectural differences. The KDA model is a hybrid with a standard final attention layer; it is not the full Kimi Linear language model architecture. The results reveal that these models adopt task mechanisms similar to those of the five non RoPE models, attesting to the considerable generality of the phenomena reported in the main text. As shown in Fig.14, all three models develop cosine like attention curves and translate their respective peaks to the appropriate offset positions to accomplish the task. Notably, kda's peak exhibits a measurable shift: because kda incorporates a one dimensional convolution that shifts the target token's information at the correct offset toward later sequence positions, the model effectively retrieves this information at a closer offset relative to the marker, resulting in a leftward displacement of the peak toward smaller offsets.

F.1.3 RESULTS ACROSS MULTIPLE SEQUENCE LENGTHS

To verify that the phenomena reported in the main text generalize across different sequence lengths, we conducted experiments in which only the sequence length was varied, with all other settings held fixed. As illustrated in the Fig.16, Fig.20, Fig.17, Fig.21, Fig.18, and Fig.22, at a sequence length of 200 and 300, all models except T5 exhibit the same phenomena as reported in the main text. The four non T5 baselines reach near perfect best and worst marker accuracy and recover smooth aligned trajectories. T5 solves some markers but retains very low worst marker accuracy and high worst marker loss (Fig.15, Fig.19).

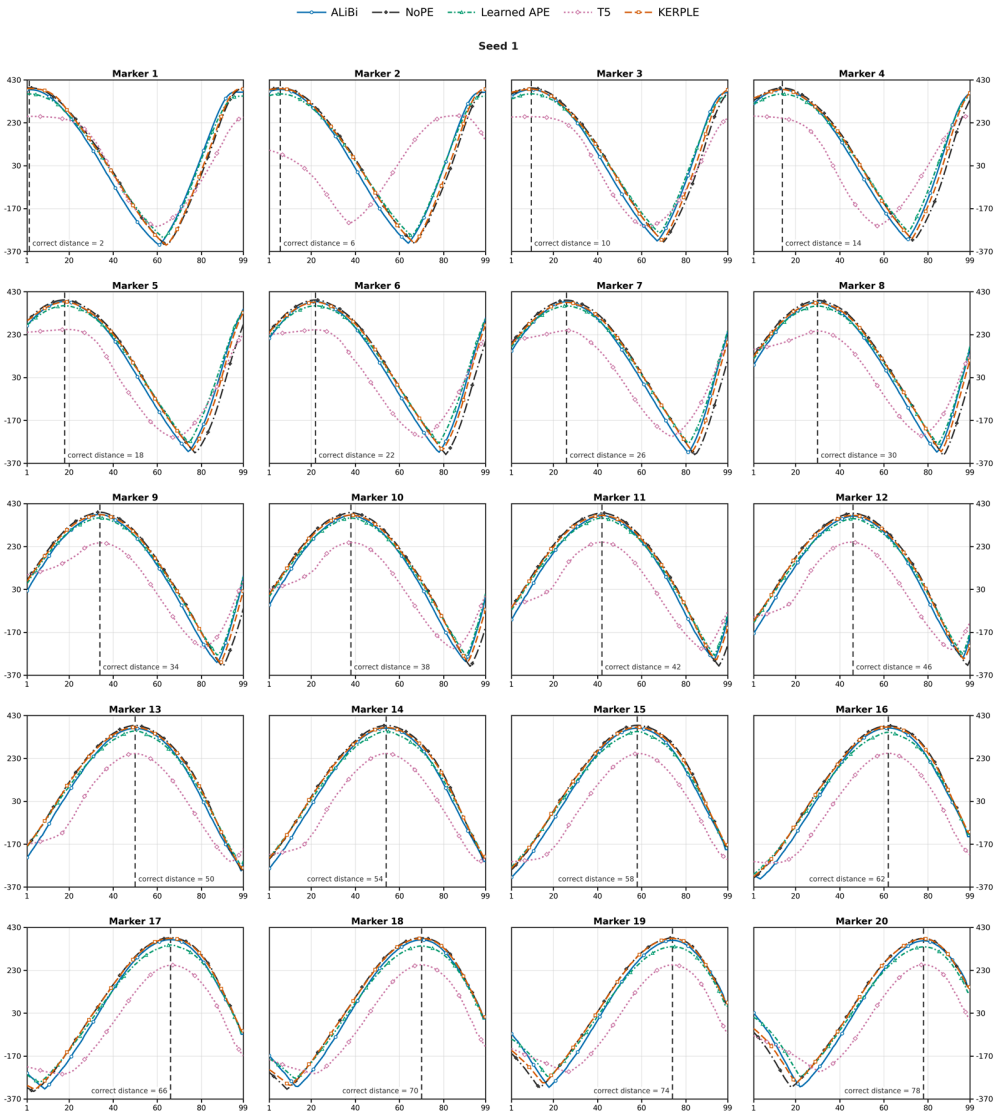

Figure 7: **QK score, seed 1.** The 5×4 grid contains markers 1–20 in row-major order for $L = 100$ at epoch 209. The horizontal coordinate is query–key distance; the vertical coordinate is the QK score reported by the source diagnostic, not a cosine similarity or a normalized attention probability. Colors and line styles identify ALiBi, NoPE, Learned APE, T5, and KERPLE. Vertical dashed lines mark the required offsets $d_m = 2 + 4(m - 1)$.

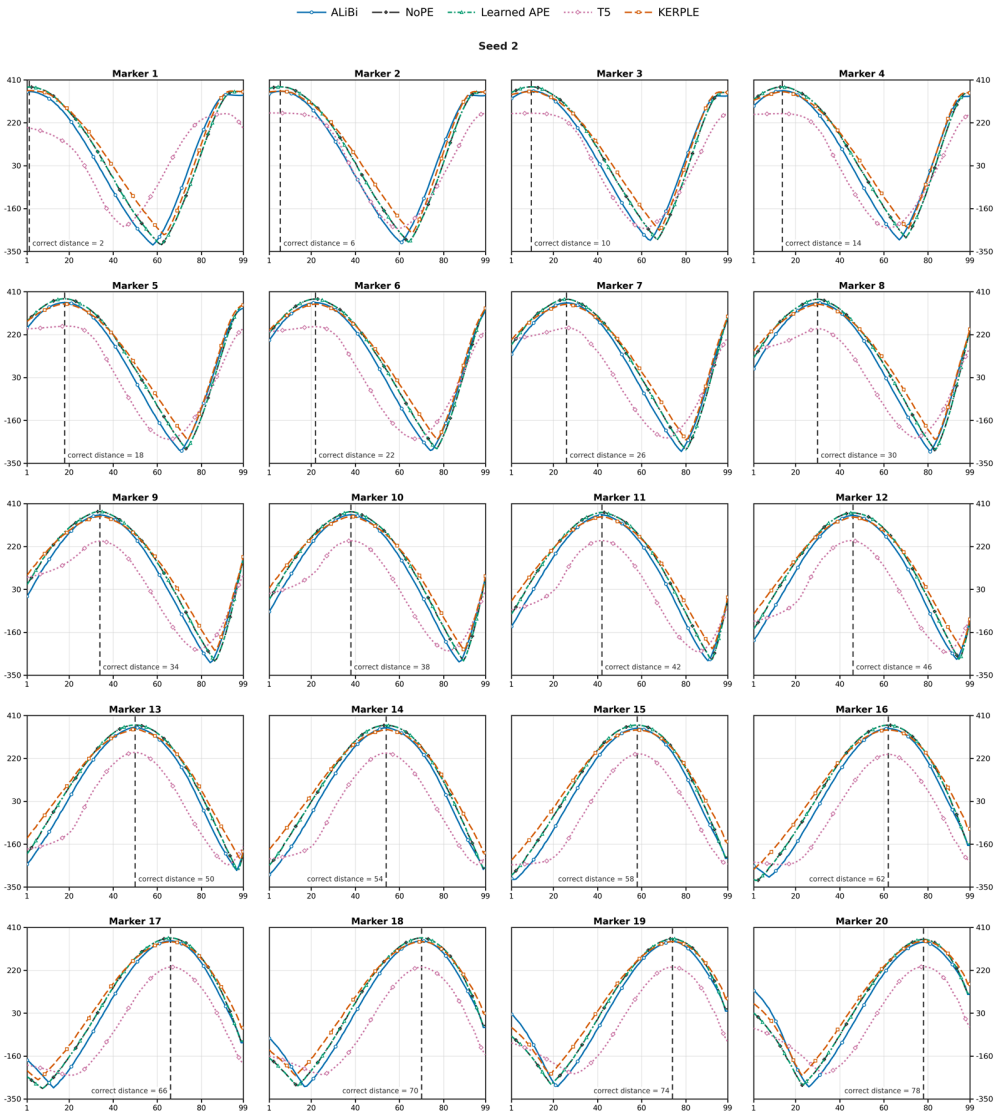

Figure 8: **QK score, seed 2.** The 5×4 grid contains markers 1–20 in row-major order for $L = 100$ at epoch 209. The horizontal coordinate is query–key distance; the vertical coordinate is the QK score reported by the source diagnostic, not a cosine similarity or a normalized attention probability. Colors and line styles identify ALiBi, NoPE, Learned APE, T5, and KERPLE. Vertical dashed lines mark the required offsets $d_m = 2 + 4(m - 1)$.

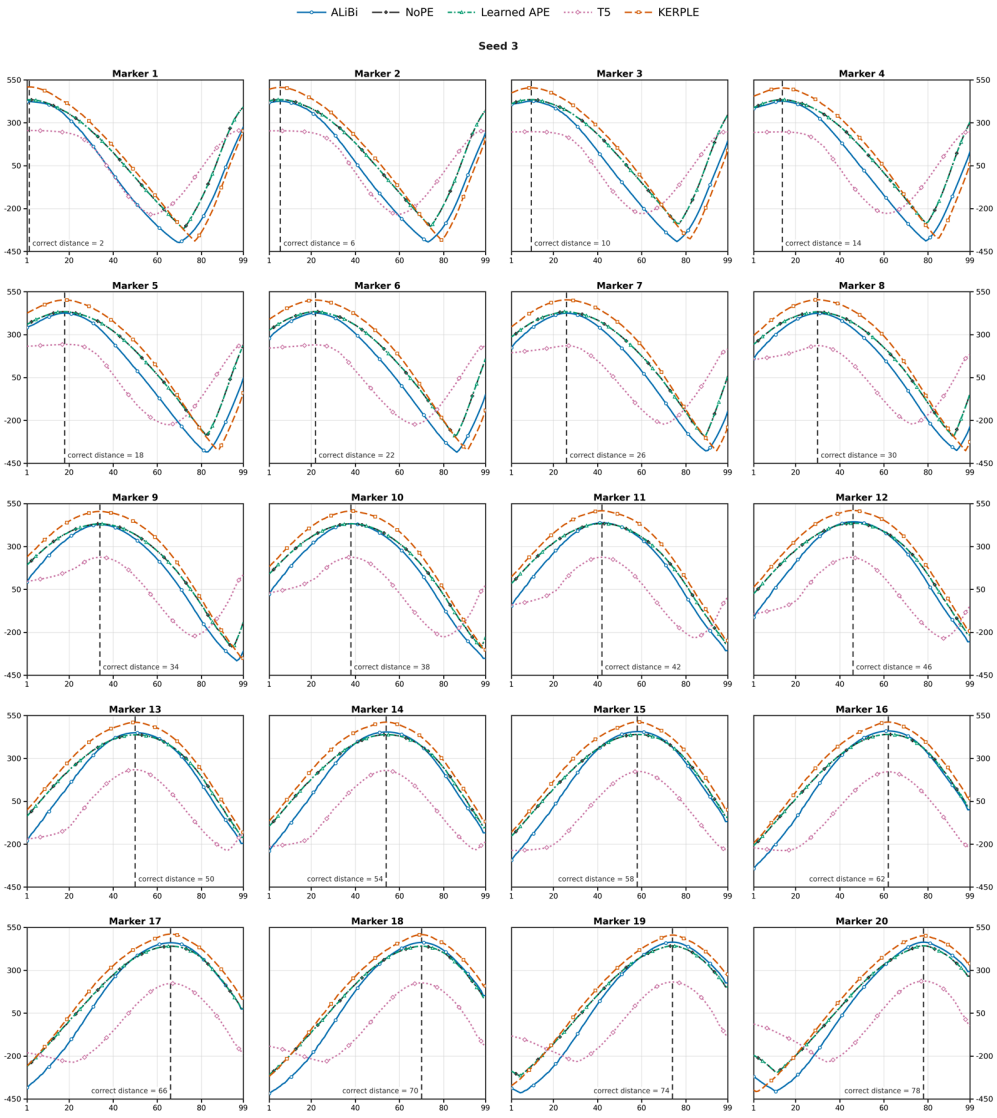

Figure 9: **QK score, seed 3.** The 5×4 grid contains markers 1–20 in row-major order for $L = 100$ at epoch 209. The horizontal coordinate is query–key distance; the vertical coordinate is the QK score reported by the source diagnostic, not a cosine similarity or a normalized attention probability. Colors and line styles identify ALiBi, NoPE, Learned APE, T5, and KERPLE. Vertical dashed lines mark the required offsets $d_m = 2 + 4(m - 1)$.

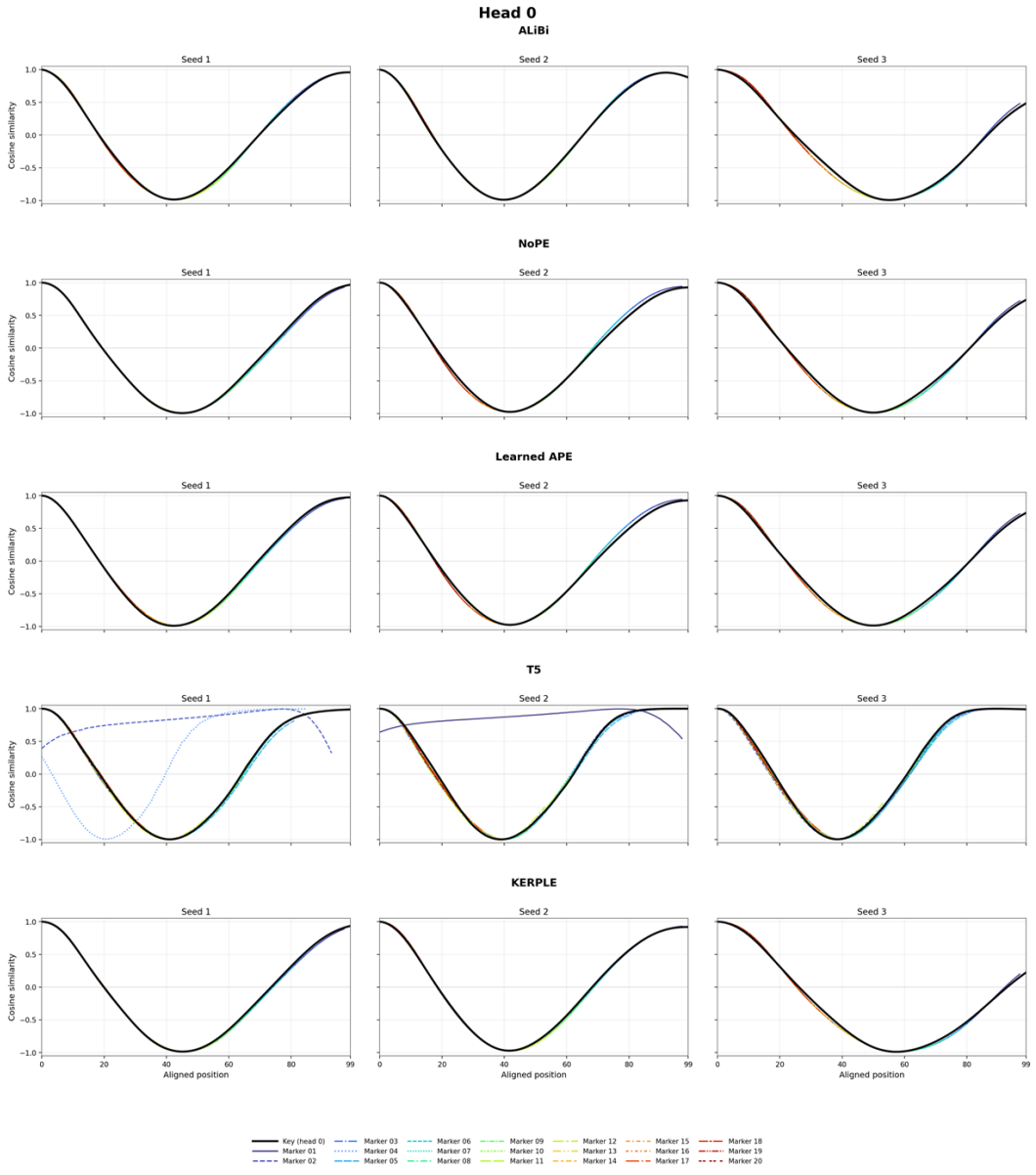

Figure 10: **Aligned key and query cosine, head 0.** Rows show the five non-RoPE encodings and columns show seeds 1–3 at $L = 100$, epoch 209, in the final layer. Black curves show the key profile; colored curves show the twenty marker-query profiles. The horizontal axis uses target-aligned coordinates: a query at position i is plotted at $i - d_m$. Each query ends at its last legal target, so different curve lengths reflect different supports rather than missing measurements. The vertical axis is cosine similarity to the key reference.

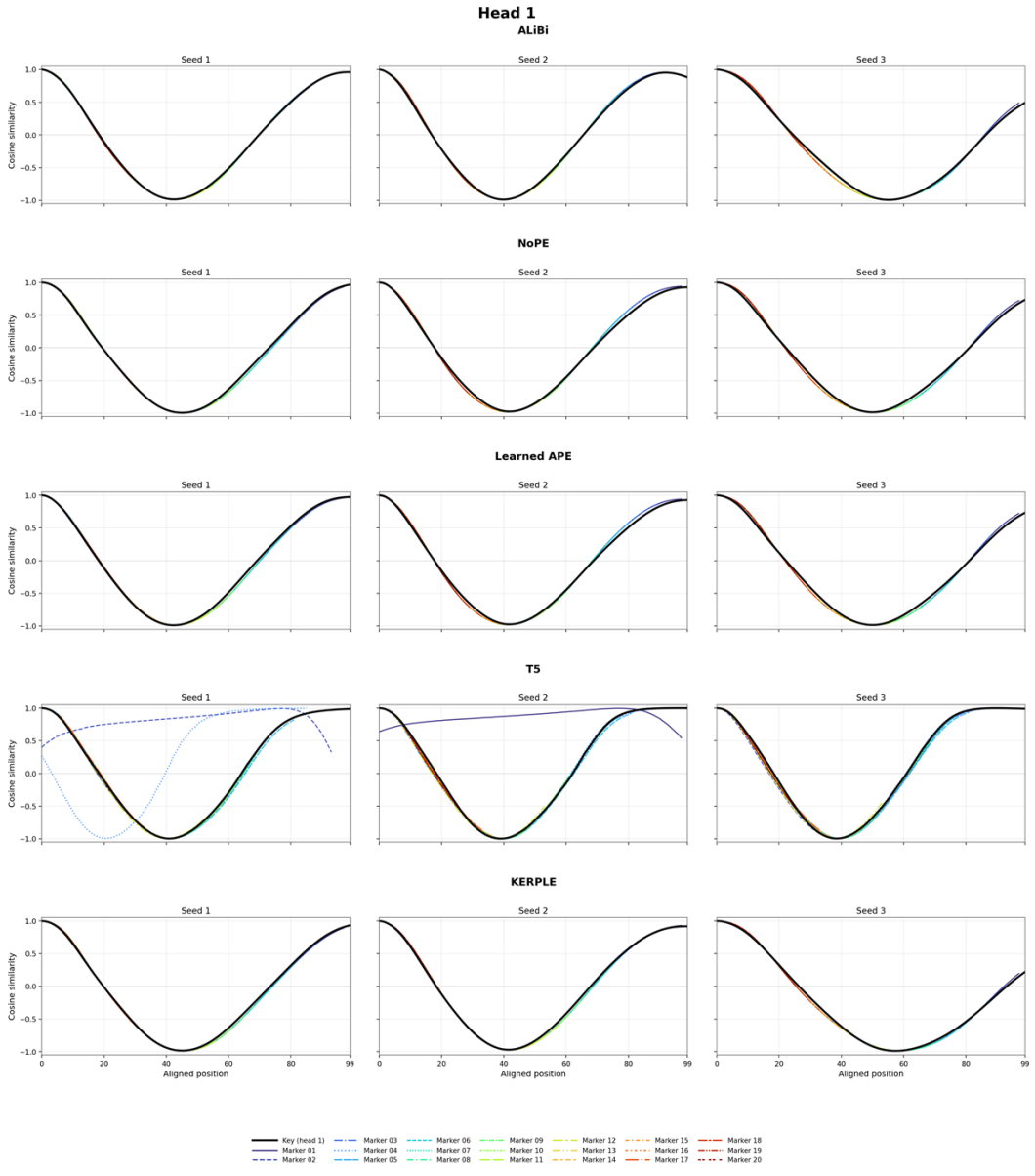

Figure 11: **Aligned key and query cosine, head 1.** Rows show the five non-RoPE encodings and columns show seeds 1–3 at $L = 100$, epoch 209, in the final layer. Black curves show the key profile; colored curves show the twenty marker-query profiles. The horizontal axis uses target-aligned coordinates: a query at position i is plotted at $i - d_m$. Each query ends at its last legal target, so different curve lengths reflect different supports rather than missing measurements. The vertical axis is cosine similarity to the key reference.

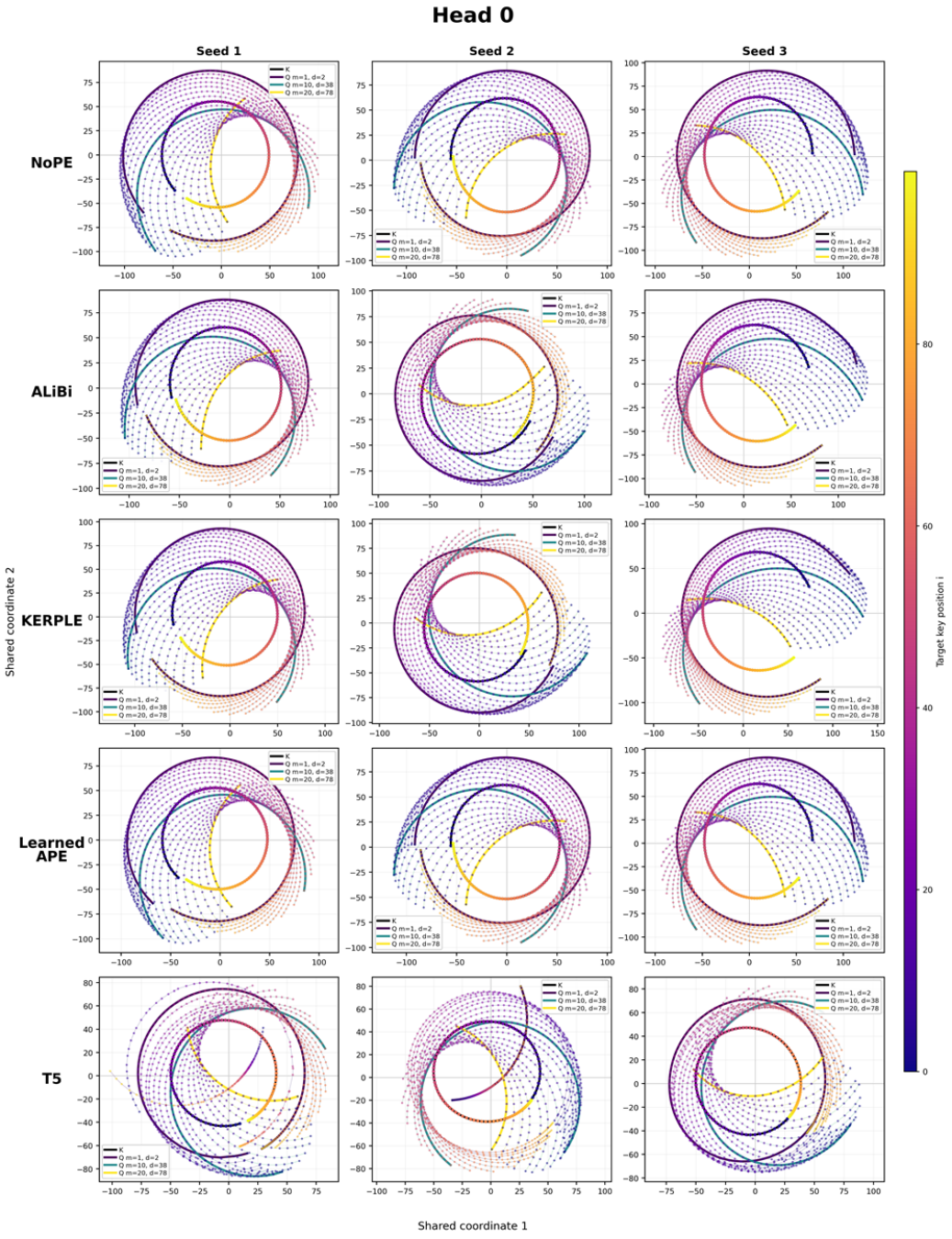

Figure 12: **Joint positional geometry, head 0.** Rows show NoPE, ALiBi, KERPLE, Learned APE, and T5 at $L = 100$. Keys and marker queries are displayed in shared two dimensional coordinates; the color bar indicates target-key position. The legend highlights the key path and markers 1, 10, and 20, with offsets 2, 38, and 78. Shorter query arcs reflect the legal supports $L - d_m$.

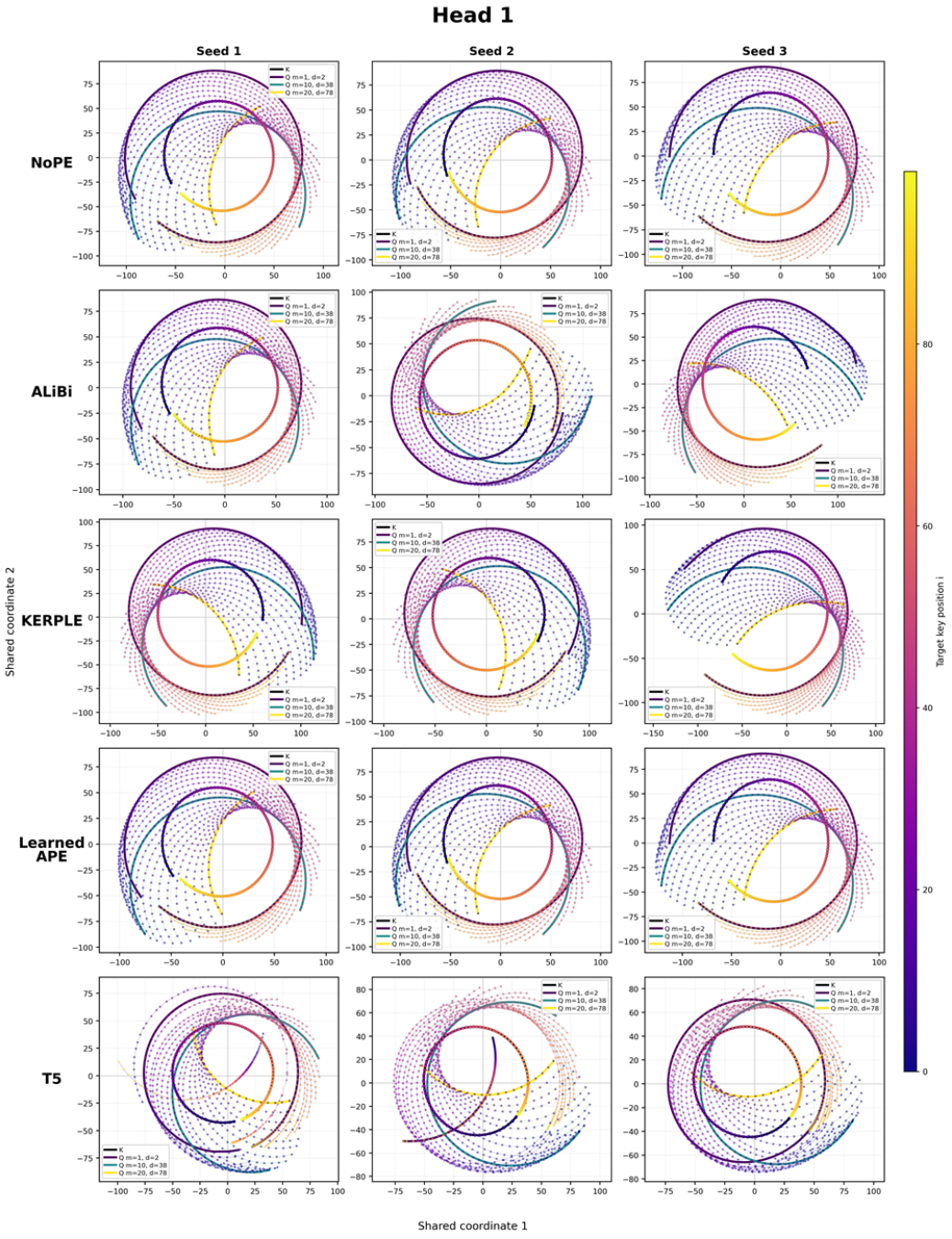

Figure 13: **Joint positional geometry, head 1.** Rows show NoPE, ALiBi, KERPLE, Learned APE, and T5 at $L = 100$. Keys and marker queries are displayed in shared two dimensional coordinates; the color bar indicates target-key position. The legend highlights the key path and markers 1, 10, and 20, with offsets 2, 38, and 78. Shorter query arcs reflect the legal supports $L - d_m$.

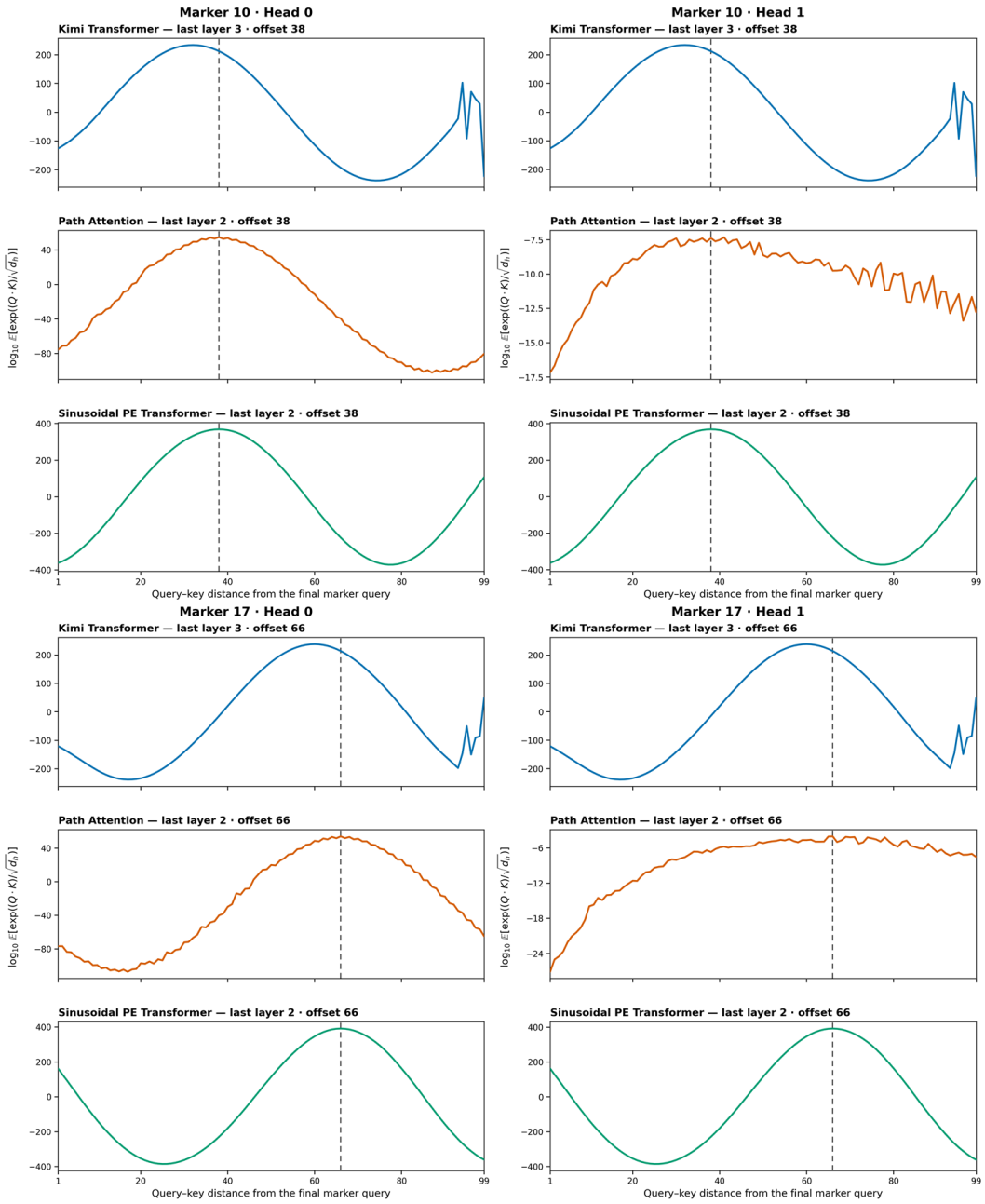

Figure 14: **QK scores in additional architectures.** The upper and lower groups show marker 10 ($d_m = 38$) and marker 17 ($d_m = 66$); columns show heads 0 and 1. Within each group, rows show the KDA hybrid, PaTH, and SinPos, using the final layers indicated in the panel titles. Horizontal axes give distance from the final marker query, and dashed lines mark the target offsets. The plotted ordinate is labeled $\log_{10} \mathbb{E}[\exp(s)]$ in the source image, with s the model's scaled pre-softmax score (including path transport for PaTH)

Table 5: **Architectures used in the additional XoM comparison.** All use model width 64, two heads, per head key width 32, feed forward width 64, and the training protocol of Table 4. Kernel sizes below describe this study's configurations.

Model	Layer layout	Positional or mixing mechanism
SinPos	Three causal Transformer layers	Fixed sine/cosine absolute embeddings with base 10,000, added to token embeddings before layer 0.
PaTH	Three causal Transformer layers	Householder path transport scores; the direction generator uses a $64 \to 32 \to 64$ bottleneck, bias free causal depthwise convolution (kernel 3), SiLU, and ℓ_2 normalization. Per head $\beta = 2\,\mathrm{sigmoid}(z) \in (0, 2)$.
KDA hybrid	Three KDA layers, then one causal Transformer layer	Causal depthwise Q/K/V convolutions (kernel 4) and Q/K ℓ_2 normalization in the KDA blocks. Diagnostics inspect the final Transformer layer (layer 3).

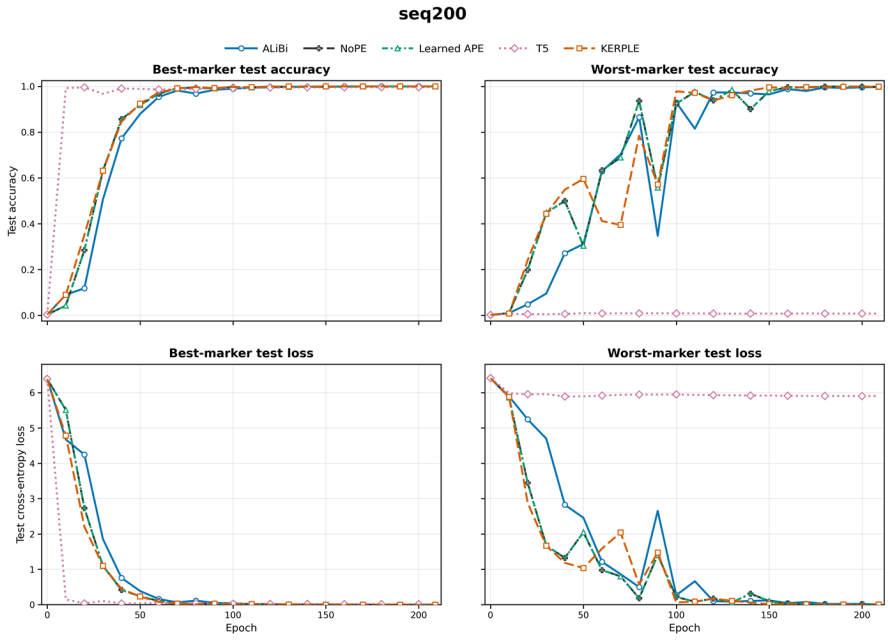

Figure 15: **Best and worst marker learning curves at $L = 200$.** The top row reports test accuracy as a fraction; the bottom row reports test cross entropy on linear vertical axes. Left panels show the best marker and right panels the worst marker according to the respective metric at each epoch. Colors identify the five non-RoPE baselines.

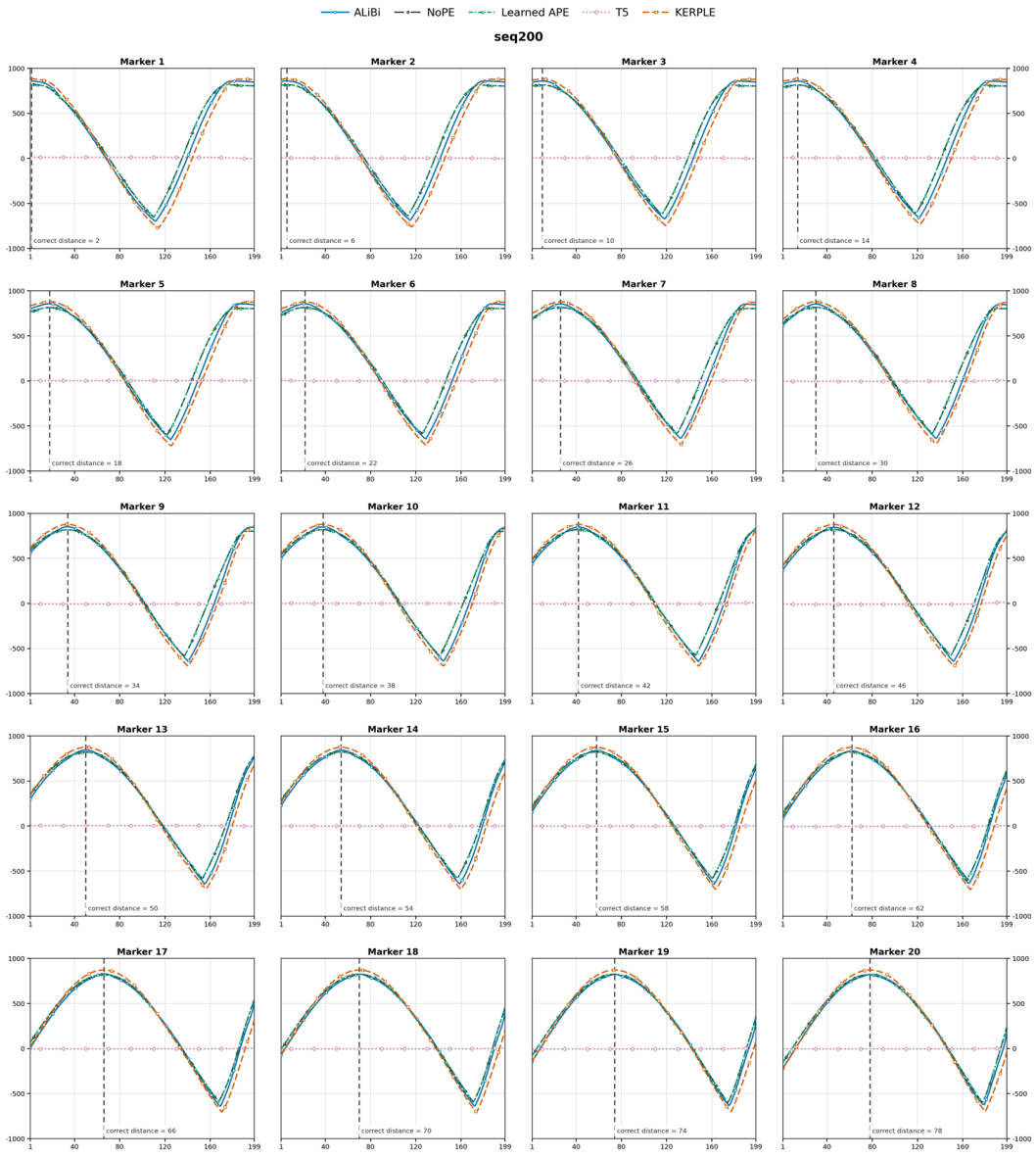

Figure 16: **QK scores at $L = 200$, final layer head 0.** Panels cover markers 1–20 in row major order at epoch 209. Horizontal coordinates are query key distances from 1 to 199, and vertical coordinates are the source QK score diagnostic. Dashed lines mark the offsets printed in the image, $d_m = 2 + 4(m - 1)$.

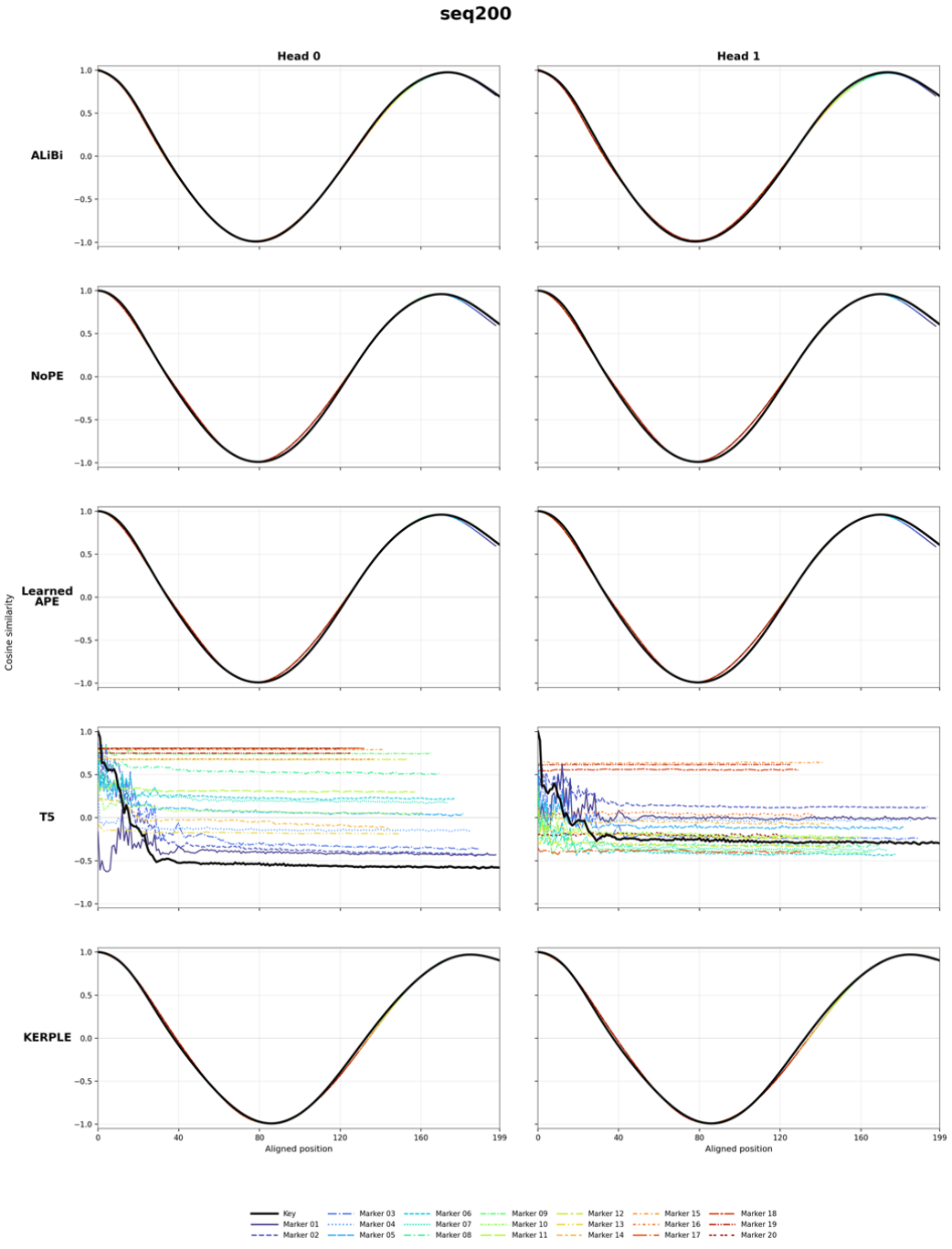

Figure 17: **Aligned key and query cosine at $L = 200$.** Rows show ALiBi, NoPE, Learned APE, T5, and KERPLE; columns show final layer heads 0 and 1 at epoch 209. Black curves are keys' and colored curves are markers'. Queries are shifted to target aligned positions $t = i - d_m$ and restricted to their legal supports.

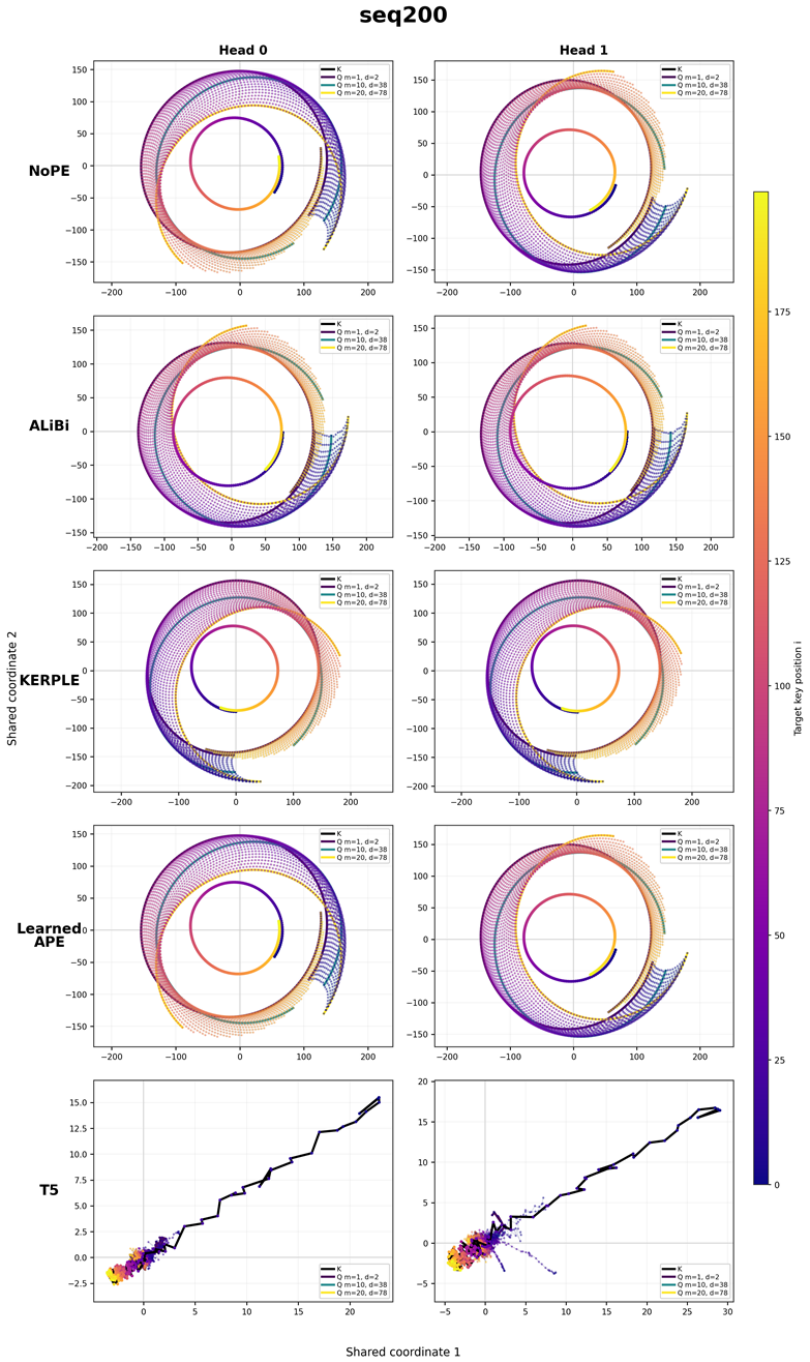

Figure 18: **Joint two dimensional positional projections at $L = 200$.** Rows show NoPE, ALiBi, KERPLE, Learned APE, and T5; columns show final layer heads 0 and 1. Color denotes target key position. The legend identifies the key path and markers 1, 10, and 20 with offsets 2, 38, and 78.

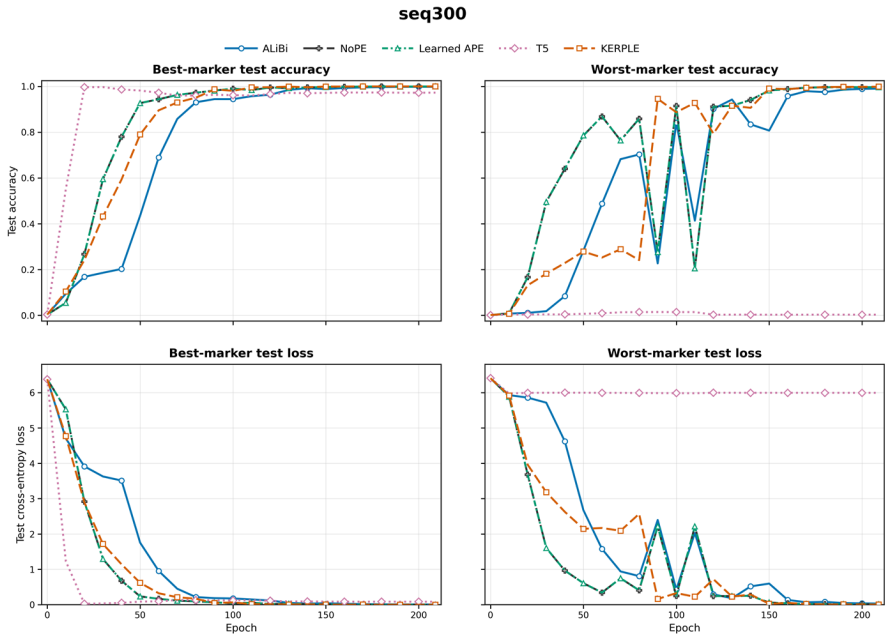

Figure 19: **Best and worst marker learning curves at $L = 300$.** The top row reports test accuracy as a fraction; the bottom row reports test cross entropy on linear vertical axes. Left panels show the best marker and right panels the worst marker according to the respective metric at each epoch. Colors identify the five non-RoPE baselines.

F.1.4 Results Across Multiple Initialization Settings

Model initialization has been established as a significant determinant of both final performance and the particular solution a model converges to. We therefore conducted additional experiments on the NoPE model across a range of initializations to investigate whether its converged solution remains consistent with the findings reported in the main text. Specifically, we varied the γ parameter across the values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. A parameter $W \in \mathbb{R}^{d_1 \times d_2}$ is initialized a Gaussian distribution $\mathcal{N}(0, (1/d_1^{\gamma})^2)$. The results reveal that, with the sole exception of $\gamma = 0.1$, at which the model fails to learn the task entirely, NoPE consistently learns a rotation like solution analogous to that described in the main text across every other initialization (Fig.23, Fig.24).

F.2 Additional In-Depth Experiments

F.2.1 Critical Head Identification Experiments

To identify the specific layer and head responsible for RoPE's task performance, We therefore examine two interventions: *attention bypass*, which sets the selected attention output to zero, and *target-value substitution*, which replaces the value at the correct source position with another sequence token's value. For RoPE, Fig.25 identify layer 1, head 1 as the pathway most sensitive to substitution of the target value. Replacing that head's target value reduces original answer accuracy to zero on the reported subset, whereas replacing layer 1, head 0 leaves it at 100%. The former intervention produces the replacement answer in 98.3250% of evaluated cases. This is direct evidence that the head carries target token information in the tested model.

Notably, these findings suggest that a single layer may be sufficient for RoPE to solve XoM, motivating our use of a single layer single head architecture in subsequent stress tests.

For the five non-RoPE models, target value substitution is most disruptive in the final layer (layer 2). Intervening on either final-layer head substantially reduces accuracy; intervening on both reduces it further (Fig.26). Thus, each model essentially employs two heads in the final layer for task completion, these are precisely the heads whose behavior is presented in the main text.

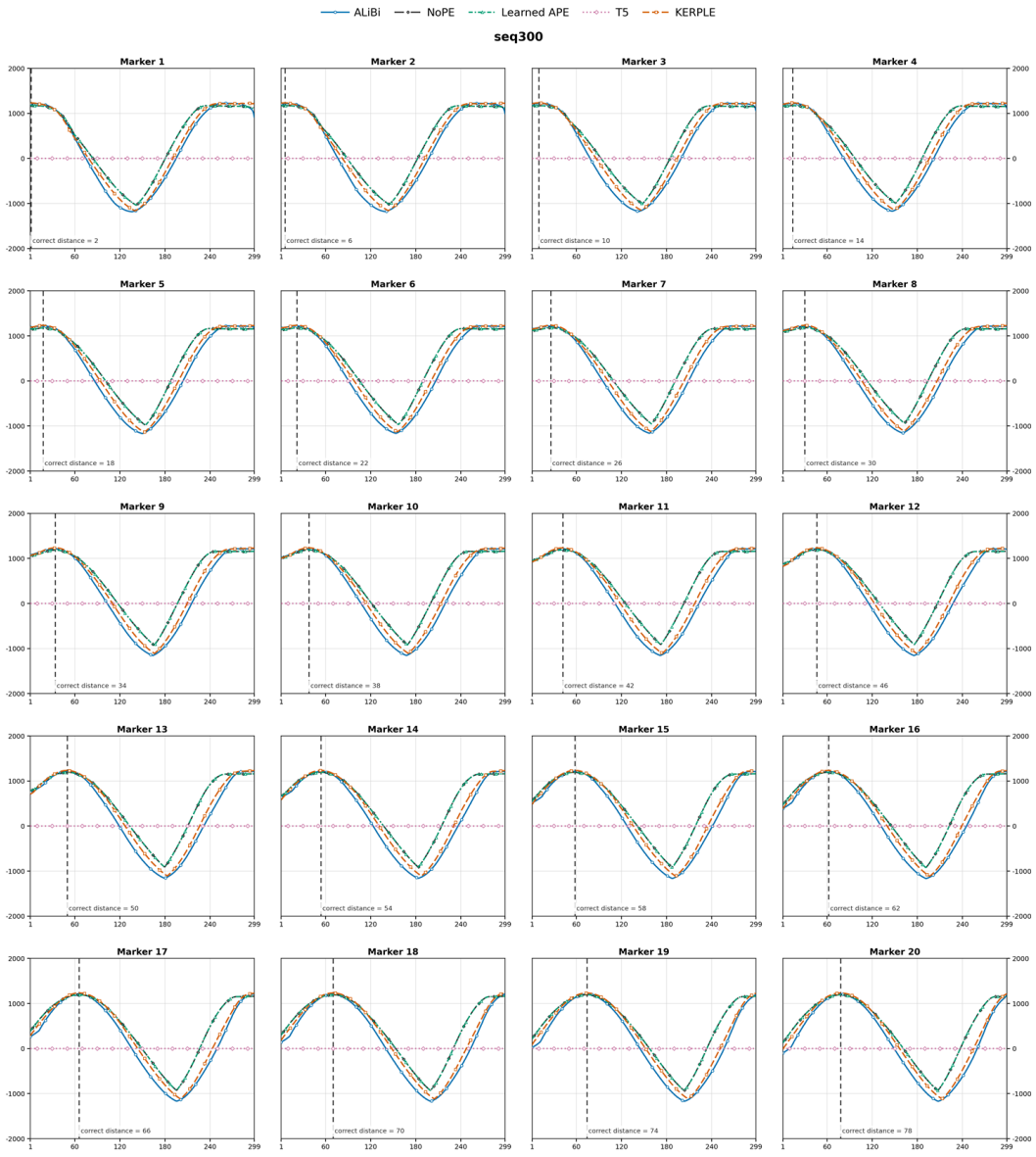

Figure 20: **QK scores at $L = 300$, final layer head 0.** Panels cover markers 1–20 in row major order at epoch 209. Horizontal coordinates are query key distances from 1 to 199, and vertical coordinates are the source QK score diagnostic. Dashed lines mark the offsets printed in the image, $d_m = 2 + 4(m - 1)$.

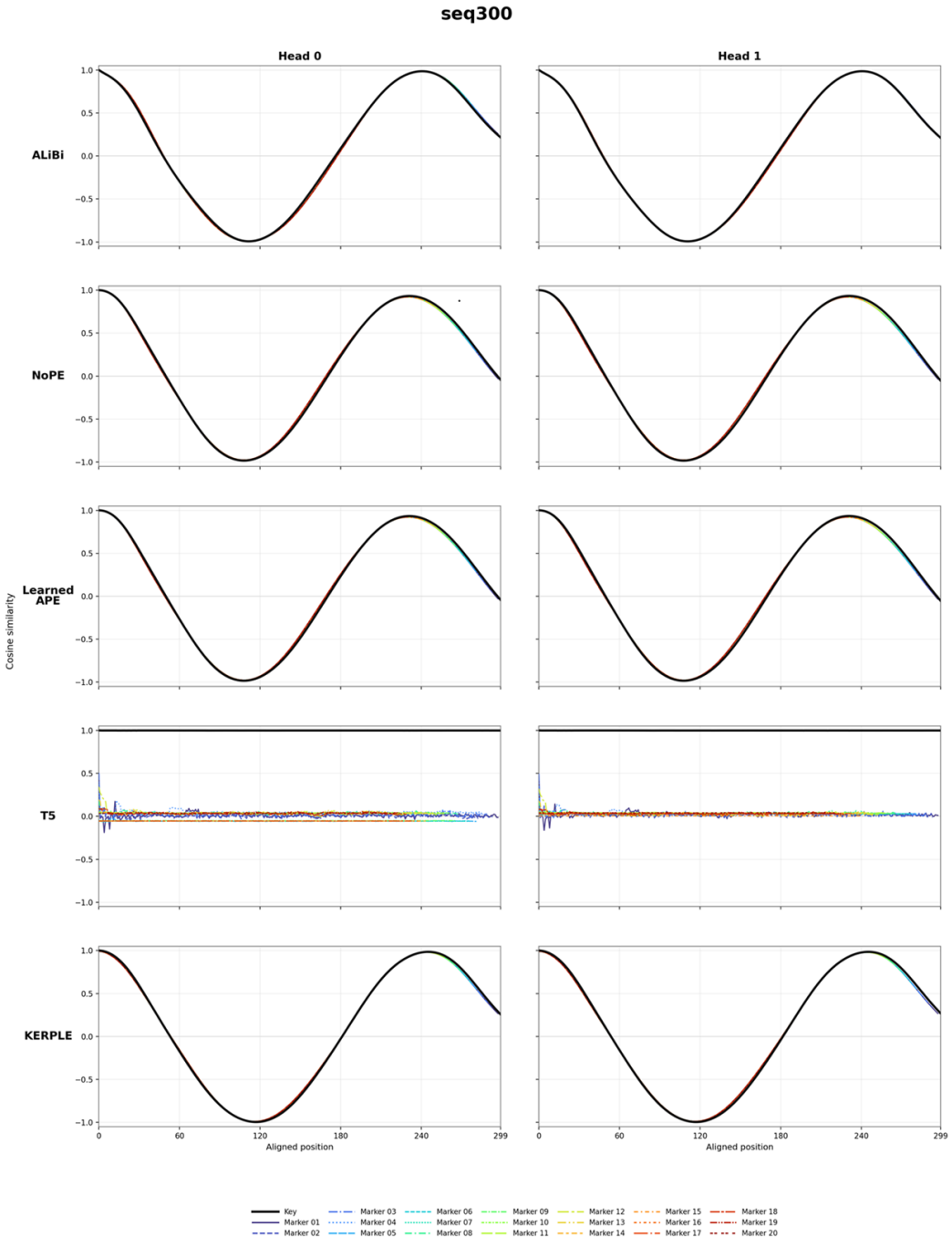

Figure 21: **Aligned key and query cosine at $L = 300$.** Rows show ALiBi, NoPE, Learned APE, T5, and KERPLE; columns show final layer heads 0 and 1 at epoch 209. Black curves are keys' and colored curves are markers'. Queries are shifted to target aligned positions $t = i - d_m$ and restricted to their legal supports.

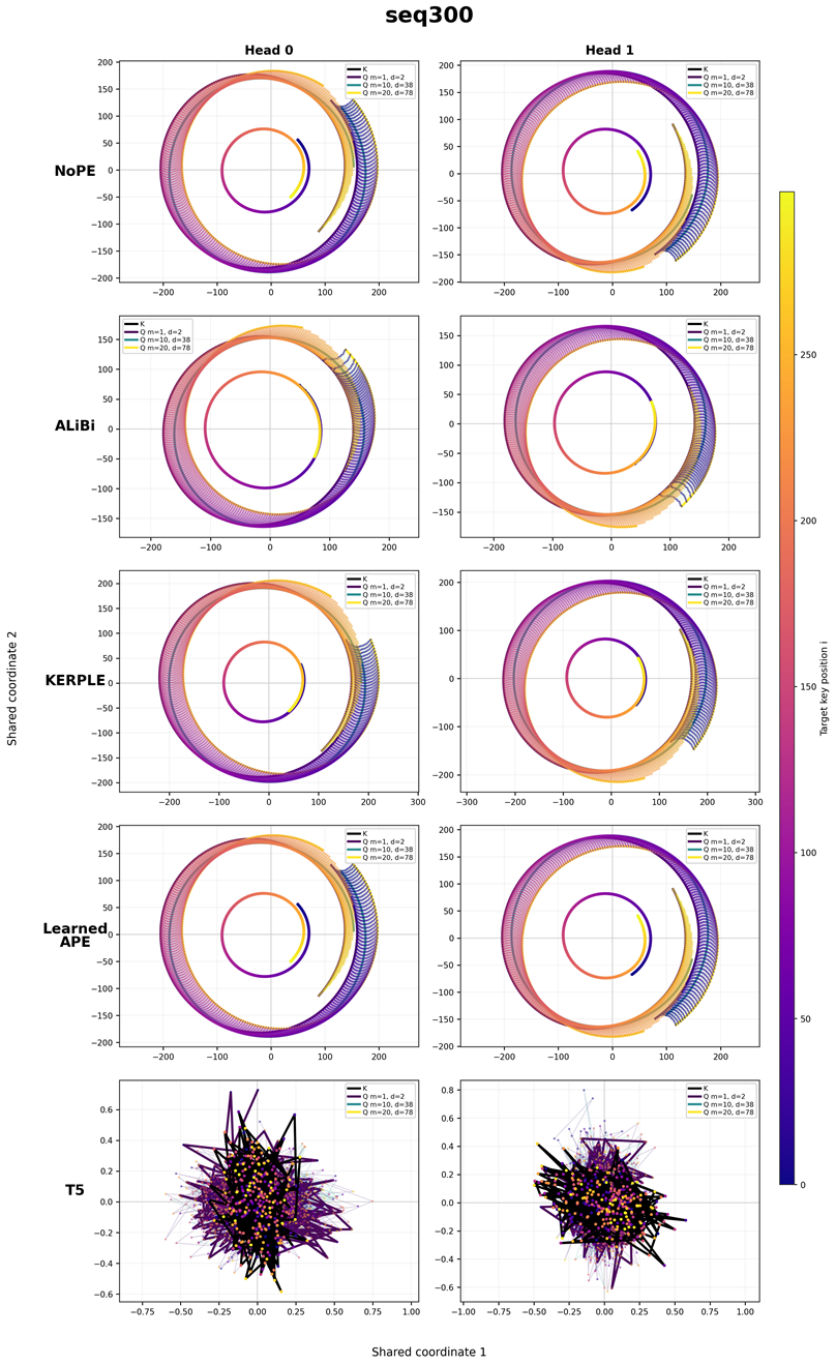

Figure 22: **Joint two dimensional positional projections at $L = 300$.** Rows show NoPE, ALiBi, KERPLE, Learned APE, and T5; columns show final layer heads 0 and 1. Color denotes target key position. The legend identifies the key path and markers 1, 10, and 20 with offsets 2, 38, and 78.

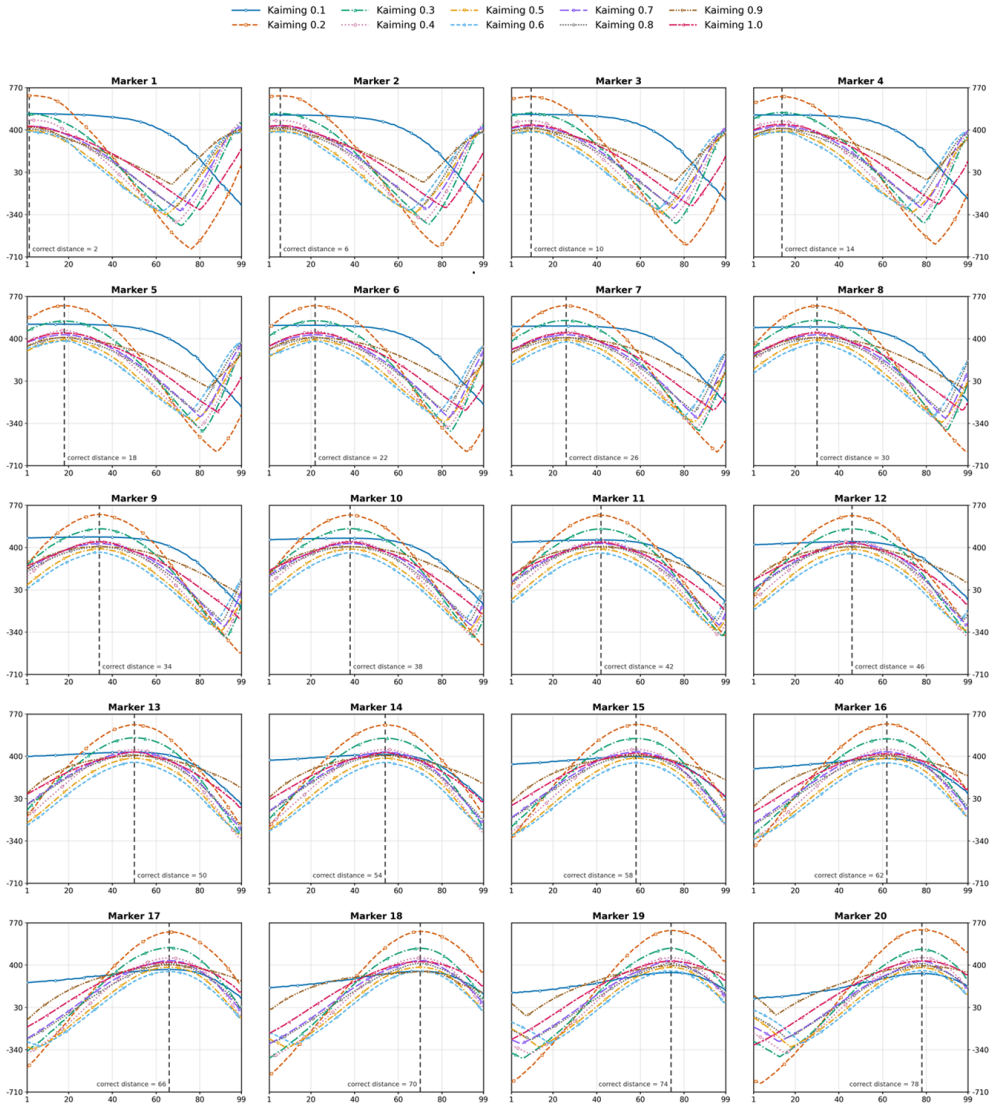
Kaiming 0.1
Kaiming 0.2
Kaiming 0.3
Kaiming 0.4
Kaiming 0.5
Kaiming 0.6
Kaiming 0.7
Kaiming 0.8
Kaiming 0.9
Kaiming 1.0
Marker 1
Marker 2
Marker 3
Marker 4
Marker 5
Marker 6
Marker 7
Marker 8
Marker 9
Marker 10
Marker 11
Marker 12
Marker 13
Marker 14
Marker 15
Marker 16
Marker 17
Marker 18
Marker 19
Marker 20

Figure 23:

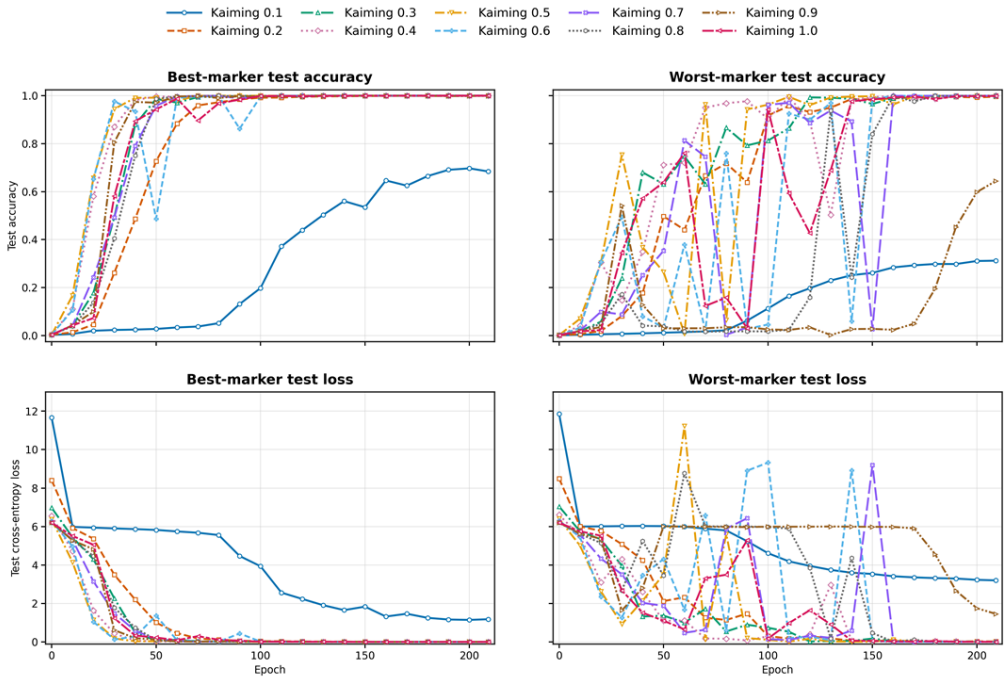

Figure 24:

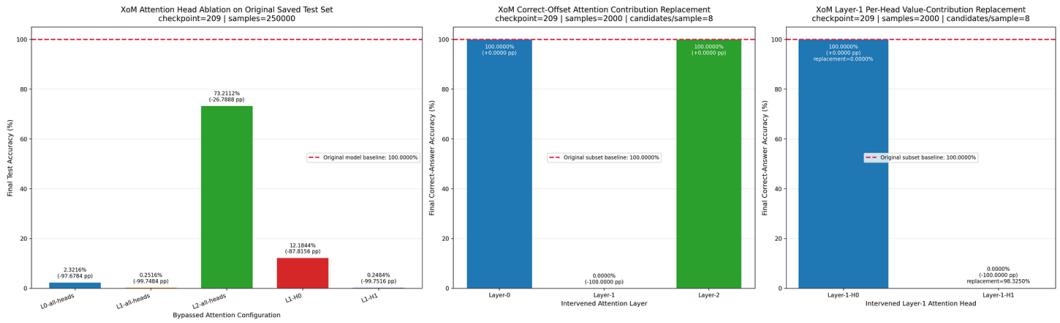

Figure 25: **RoPE head identification.** Left: accuracy after bypassing the indicated attention layers or heads on 250,000 saved test sequences. Center: target value substitution by layer on 2,000 sequences, with eight replacement candidates per sequence. Right: the same substitution restricted to one head of layer 1. Red dashed lines show the original 100% baselines; bar annotations report accuracy and its change in percentage points. The right panel also reports replacement-answer rates.

F.2.2 Angular Trajectories of Key and Query in RoPE

To demonstrate that the angular variation of the marker's query with absolute position matches that of the key, we constructed separate cosine trajectory plots for the query and key. Figure 27 compares key and query cosine profiles against a common reference key at position 0. The displayed mean averages over 128 reference tokens. After shifting the query coordinate by its marker offset, the curves show closely corresponding oscillatory structure. This reveals that the angular changes of the marker's query and key occur in lockstep.

F.2.3 Experiments on Norm

To confirm that the model solves XoM by virtue of the angular variations of the key and the marker's query. We plot the key norm as a function of absolute position, alongside the query norm for each

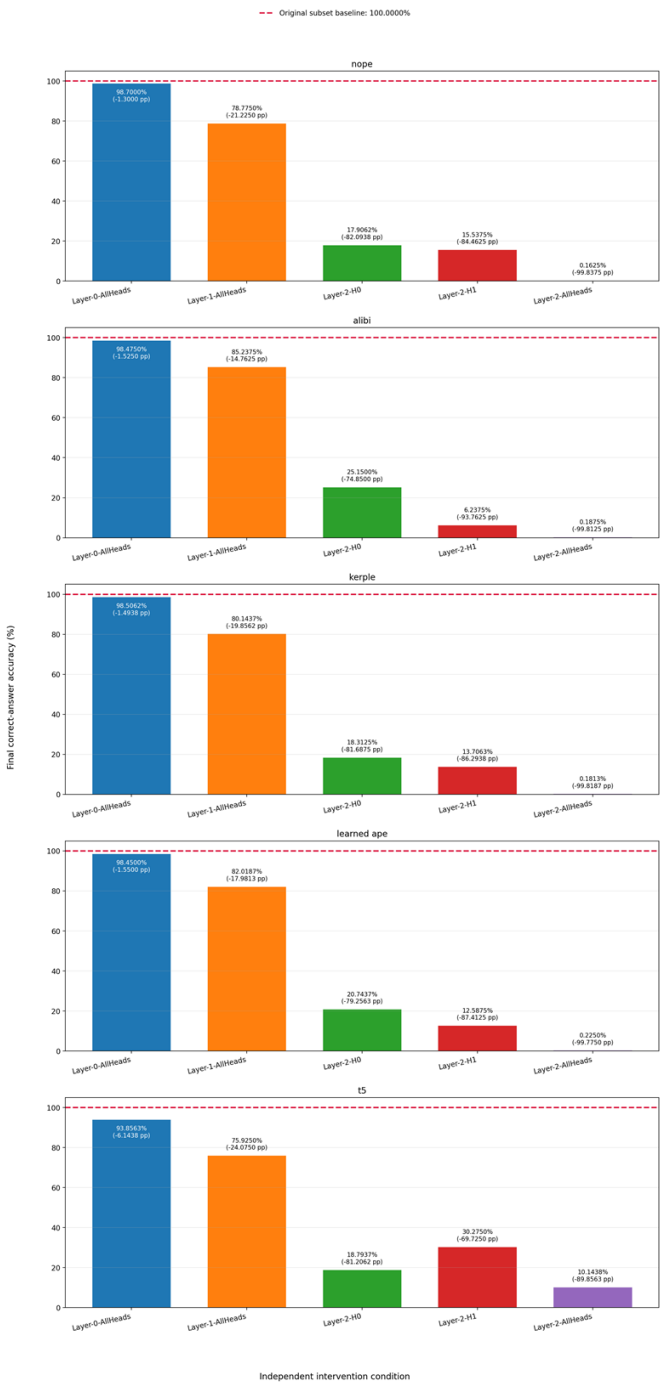

Figure 26: **Target value substitution in non-RoPE models.** Rows show NoPE, ALiBi, KERPLE, Learned APE, and T5. Within each row, the five bars intervene on both heads of layer 0, both heads of layer 1, head 0 of layer 2, head 1 of layer 2, and both heads of layer 2, respectively. Vertical axes report accuracy, red dashed lines mark the 100% subset baseline. Bar annotations give accuracy and change in percentage points. Each condition is evaluated independently.

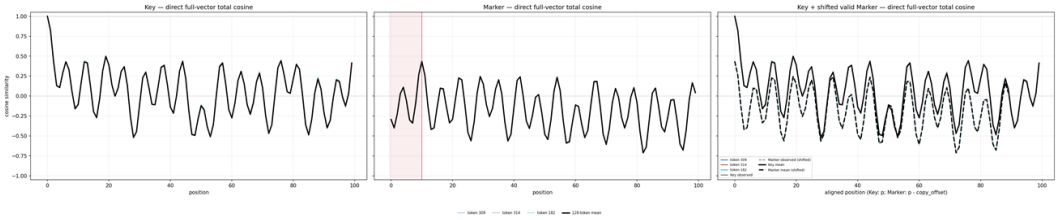

Figure 27: **Angular curves in RoPE.** Left: cosine similarity between a token's key at position p and its reference key at position 0. Center: the marker query's cosine similarity to the same reference key, indexed by physical marker position; the shaded prefix precedes the legal offset boundary. Right: key and query profiles overlaid after shifting query positions by the copy offset and retaining only valid queries. Thin colored curves illustrate reference tokens and thick curves show the 128-token mean; solid and dashed lines distinguish keys and shifted queries.

marker across positions. Fig.28, Fig.29, and Fig.30 show key norms for seeds 1–3; Fig.31, Fig.32, and Fig.33 show query norms for markers 1, 10, and 20. Key and query norms vary with position and across seeds, while their variation is generally modest relative to the overall norm. These plots support approximate radial stability in the displayed conditions, rather than exactly constant norms. Norm variation of this kind is incapable of supporting translation invariant relative position retrieval.

F.2.4 Experiments on spectrum

In the main text, we showed that key energy concentrates in the frequency band 1. In fact, the query follows a similar pattern. We extend the main-text key analysis to all twenty marker queries, both final layer heads, and three seeds (Fig.34,Fig.35,Fig. 36). The queries of most markers concentrate their spectral energy in the first frequency band.

F.3 Comparative Experiments of RoPE and Geometric RoPE

F.3.1 Comparison Under XoM

To thoroughly compare RoPE and Geometric RoPE on XoM, we conducted experiments across a range of sequence lengths. To maximize the visibility of their differences, we minimized the model size; building on the earlier finding that RoPE can likely solve the task with a single layer in a single head, we reduced the model to one layer. This reduces the capacity for additional layers to compensate for weak positional separation. All other settings remained unchanged as Table.4. Fig.37 shows a larger minimum normalized QK margin for Geometric RoPE in every pair. The comparison uses the unit normalized, maximum distractor margin in Equation (E.4). Fig.38 shows that fewer epochs (corresponding to less data, with 2 million samples per epoch) are required in Geometric rope to reach a test loss on the order of 10^{-8}; this advantage becomes more pronounced with increasing sequence length.

F.3.2 60M Model Experiments

Both the 60M Geometric rope and RoPE control use OLMo (8 layers, width 512, 8 attention heads, 64 dimensions per head, approximately 59.32M non embedding parameters). Starting from the same initial weights, they are trained for 1 epoch on the same C4 corpus with length 512, global batch 1024, BF16, AdamW (learning rate 0.0006, weight decay 0.1, cosine decay after 10,000 warmup steps), seed 6198, and the models at step 78,019 are compared. Only at layers 3 and 6, all heads use fixed Geometric RoPE with Q=32, h=2, period 1023, no zero frequency; all other layers and the control use standard RoPE with $base = 10000$, and rotations are applied to Q/K. The two arms use 8 GPUs. For the absolute position probe, the main model is frozen, and non overlapping 1024/128/256 sequences of length 512 are taken from the same C4 validation corpus as probe training/validation/test sets. Unbiased $512 \to 1024 \to 512$ ReLU classifiers are trained on the embedding and each layer's raw residual representation respectively, predicting positions 0–511 with cross entropy. The probe uses AdamW (learning rate 0.002, weight decay 0.01), samples 8192 tokens per step, trains for 10,000 steps, cosine decay after 500 warmup steps, selects best weights

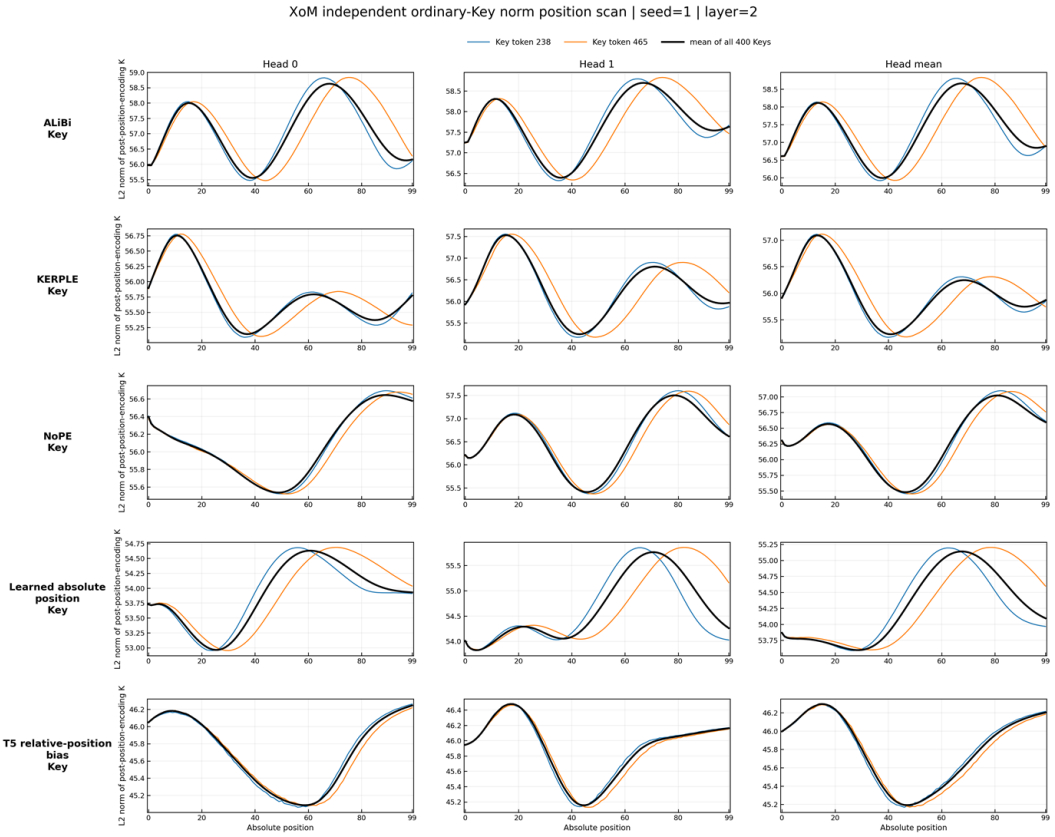

Figure 28: **Key norms seed 1.** At $L = 100$ and layer 2, rows show ALiBi, KERPLE, NoPE, Learned APE, and T5; columns show head 0, head 1, and the mean across heads. Vertical axes give the ℓ_2 norm of the key, and the horizontal axis gives absolute position. Blue and orange curves correspond to token IDs 238 and 465; the black curve is the mean of the norms over all 400 ordinary token IDs.

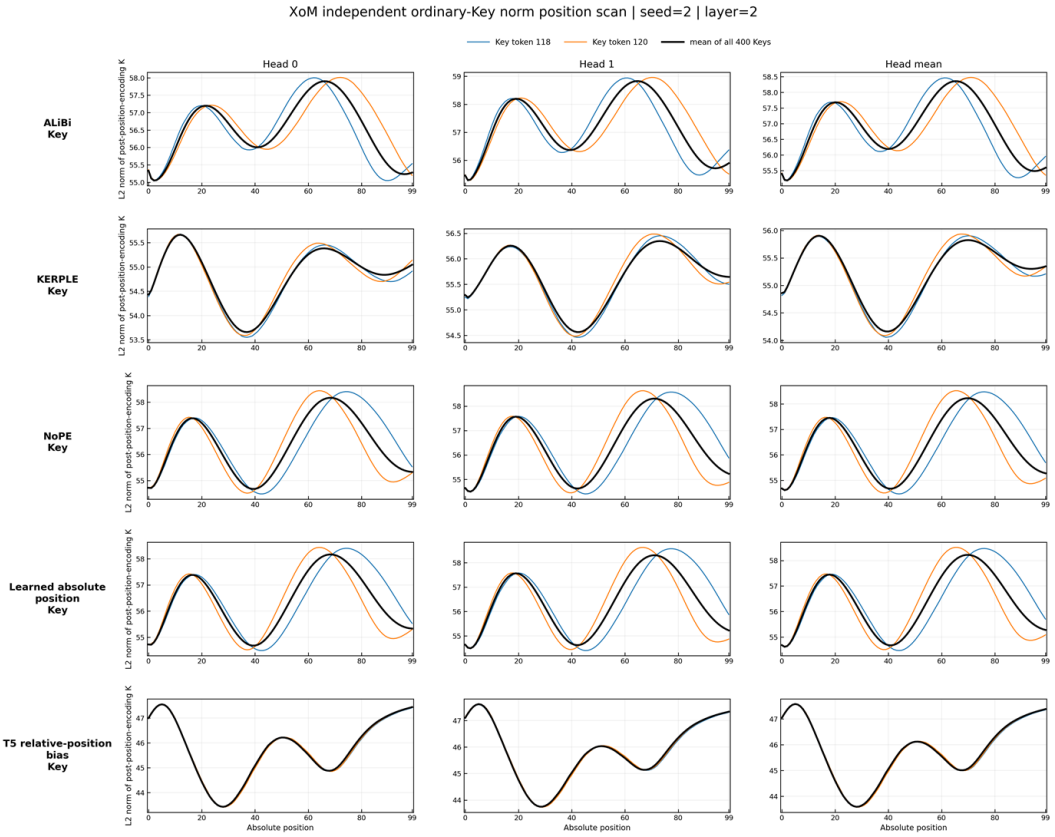

Figure 29: **Key norms seed 2.** At $L = 100$ and layer 2, rows show ALiBi, KERPLE, NoPE, Learned APE, and T5; columns show head 0, head 1, and the mean across heads. Vertical axes give the ℓ_2 norm of the key, and the horizontal axis gives absolute position. Blue and orange curves correspond to token IDs 238 and 465; the black curve is the mean of the norms over all 400 ordinary token IDs.

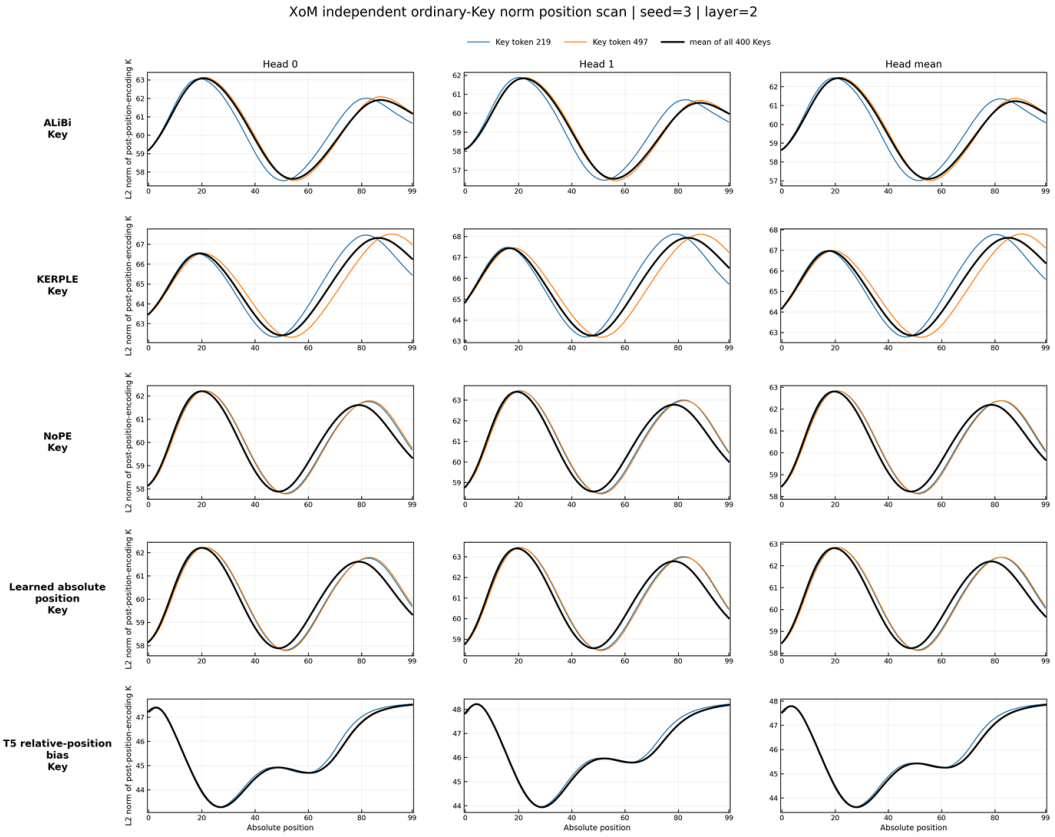

Figure 30: **Key norms seed 3.** At $L = 100$ and layer 2, rows show ALiBi, KERPLE, NoPE, Learned APE, and T5; columns show head 0, head 1, and the mean across heads. Vertical axes give the ℓ_2 norm of the key, and the horizontal axis gives absolute position. Blue and orange curves correspond to token IDs 238 and 465; the black curve is the mean of the norms over all 400 ordinary token IDs.

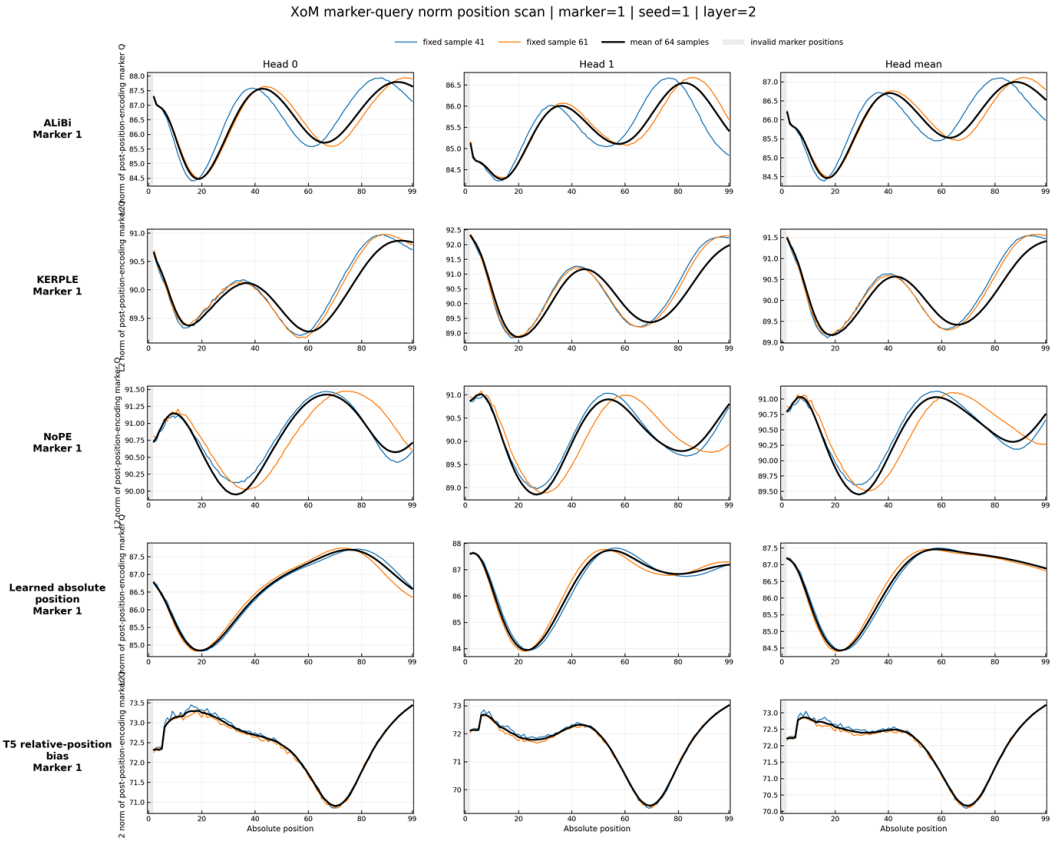

Figure 31: **Query norms for marker 1 ($d_m = 2$), seed 1.** At $L = 100$ and layer 2, rows show the five non-RoPE models in the order indicated, and columns show head 0, head 1, and the mean across heads. Horizontal axes give physical marker position; vertical axes give query ℓ_2 norms. Blue and orange curves show fixed background samples 41 and 61, and black curves average the norms over 64 samples. Gray shading marks invalid positions $i < d_m$; these positions are excluded from the interpretation.

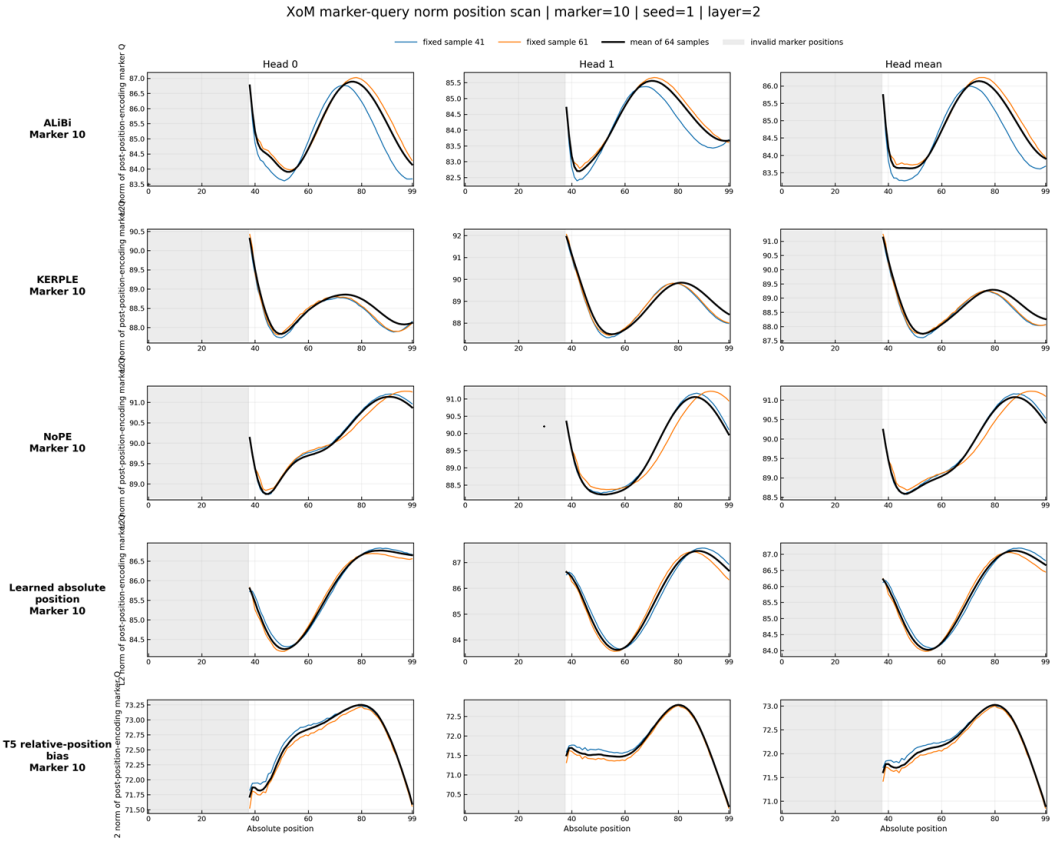

Figure 32: **Query norms for marker 10 ($d_m = 38$), seed 1.** At $L = 100$ and layer 2, rows show the five non-RoPE models in the order indicated, and columns show head 0, head 1, and the mean across heads. Horizontal axes give physical marker position; vertical axes give query ℓ_2 norms. Blue and orange curves show fixed background samples 41 and 61, and black curves average the norms over 64 samples. Gray shading marks invalid positions $i < d_m$; these positions are excluded from the interpretation.

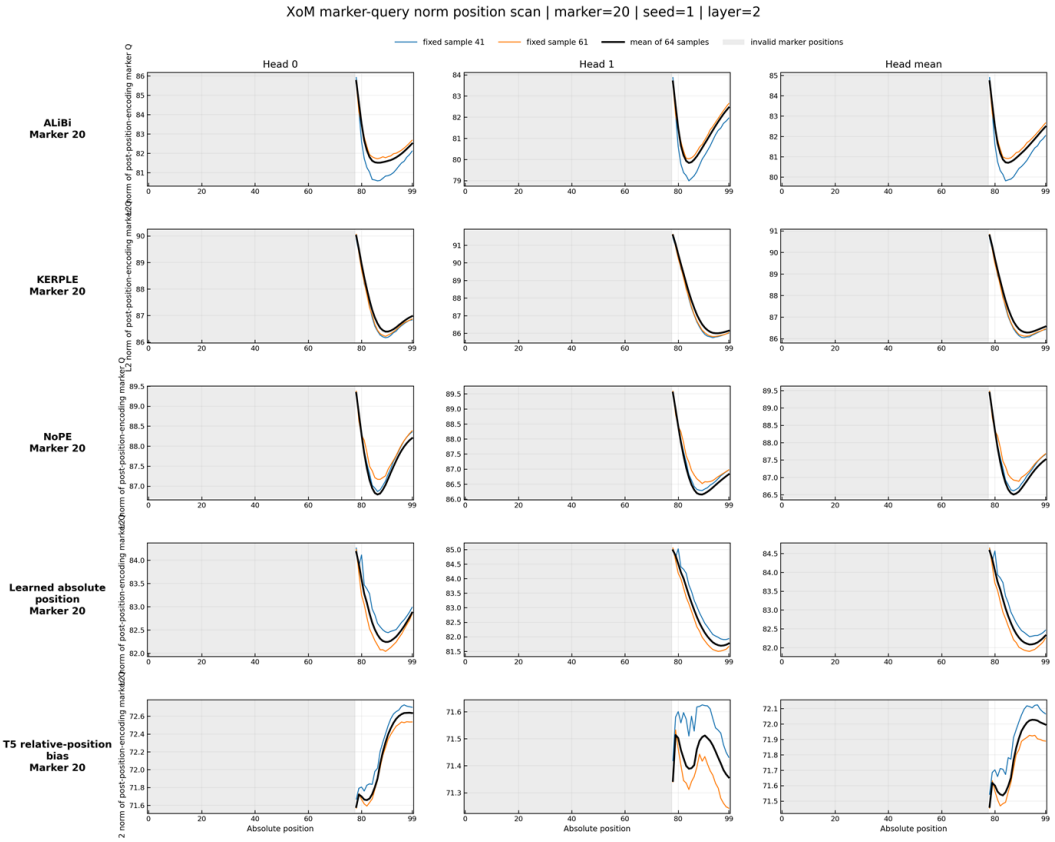

Figure 33: **Query norms for marker 20 ($d_m = 78$), seed 1.** At $L = 100$ and layer 2, rows show the five non-RoPE models in the order indicated, and columns show head 0, head 1, and the mean across heads. Horizontal axes give physical marker position; vertical axes give query ℓ_2 norms. Blue and orange curves show fixed background samples 41 and 61, and black curves average the norms over 64 samples. Gray shading marks invalid positions $i < d_m$; these positions are excluded from the interpretation.

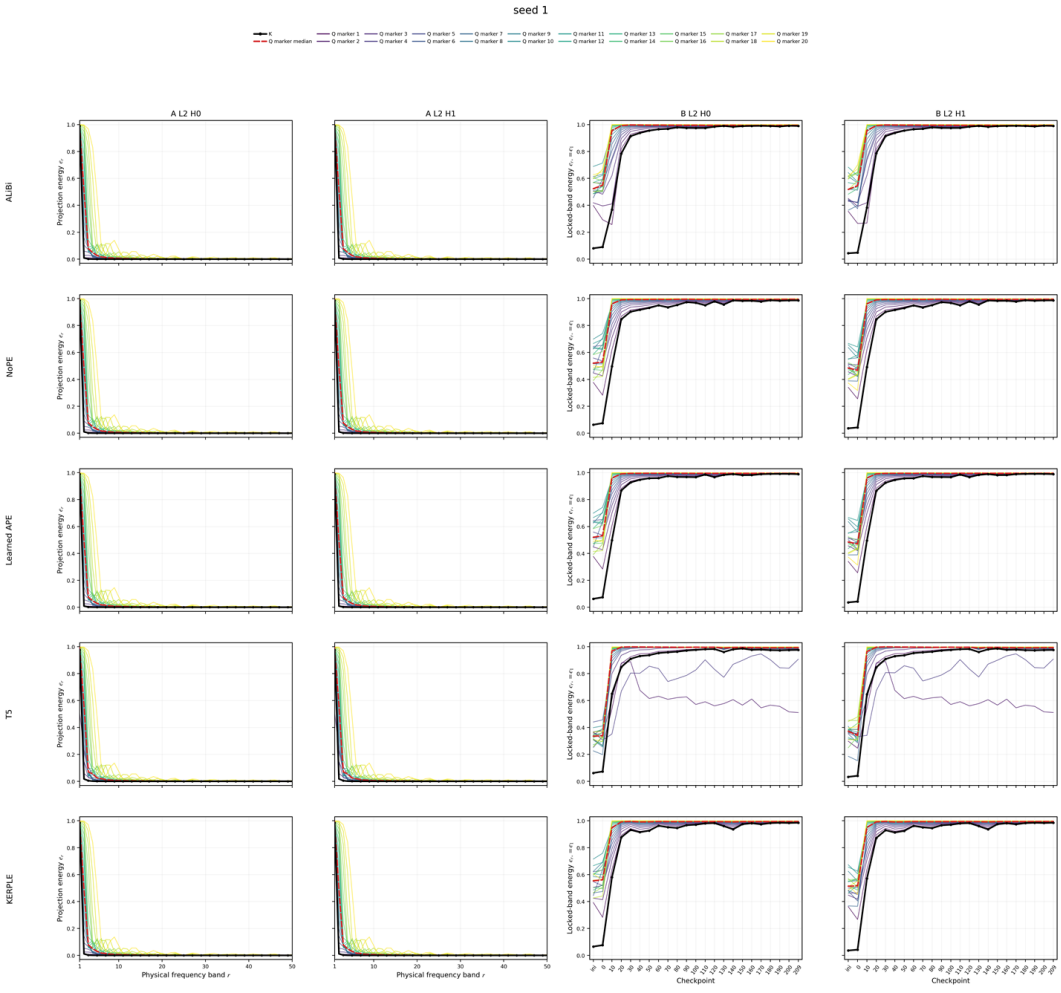

Figure 34: **Frequency fits and training dynamics seed 1.** Rows show ALiBi, NoPE, Learned APE, T5, and KERPLE. The two left columns (A) plot the final checkpoint projection score e_r against physical band index r for layer 2 heads 0 and 1. The two right columns (B) track the first band score e_1 from initialization through epoch 209. Black curves denote keys, colored curves denote individual marker queries, and the red dashed curve is the median over marker queries.

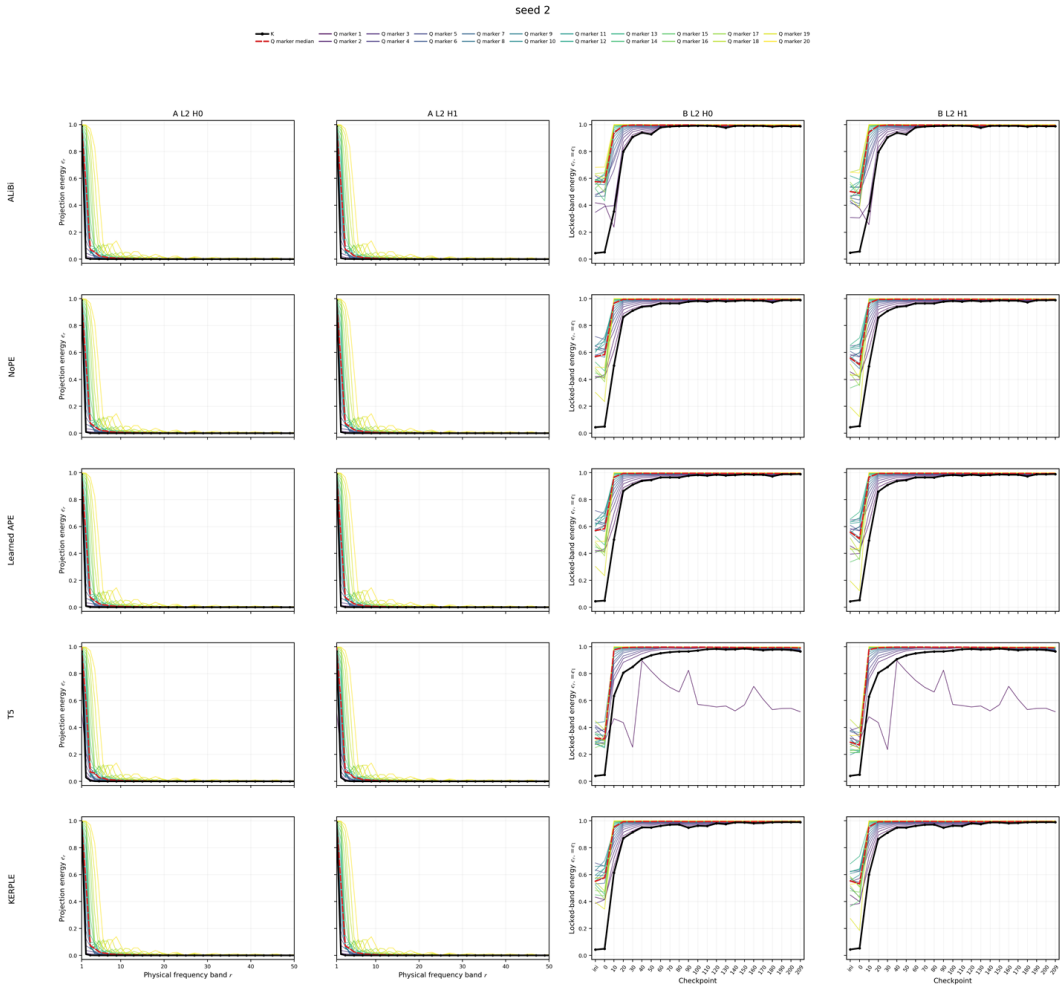

Figure 35: **Frequency fits and training dynamics seed 2.** Rows show ALiBi, NoPE, Learned APE, T5, and KERPLE. The two left columns (A) plot the final checkpoint projection score e_r against physical band index r for layer 2 heads 0 and 1. The two right columns (B) track the first band score e_1 from initialization through epoch 209. Black curves denote keys, colored curves denote individual marker queries, and the red dashed curve is the median over marker queries.

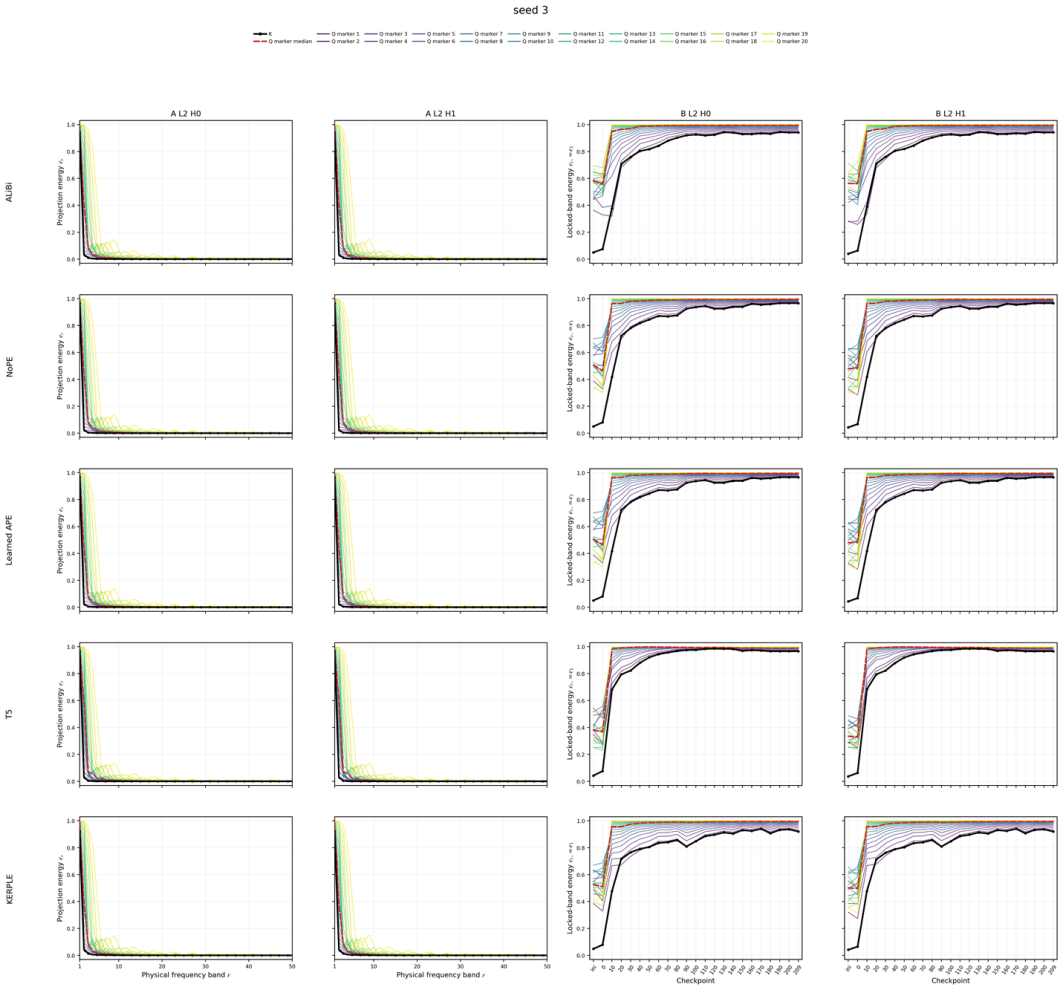

Figure 36: **Frequency fits and training dynamics seed 3.** Rows show ALiBi, NoPE, Learned APE, T5, and KERPLE. The two left columns (A) plot the final checkpoint projection score e_r against physical band index r for layer 2 heads 0 and 1. The two right columns (B) track the first band score e_1 from initialization through epoch 209. Black curves denote keys, colored curves denote individual marker queries, and the red dashed curve is the median over marker queries.

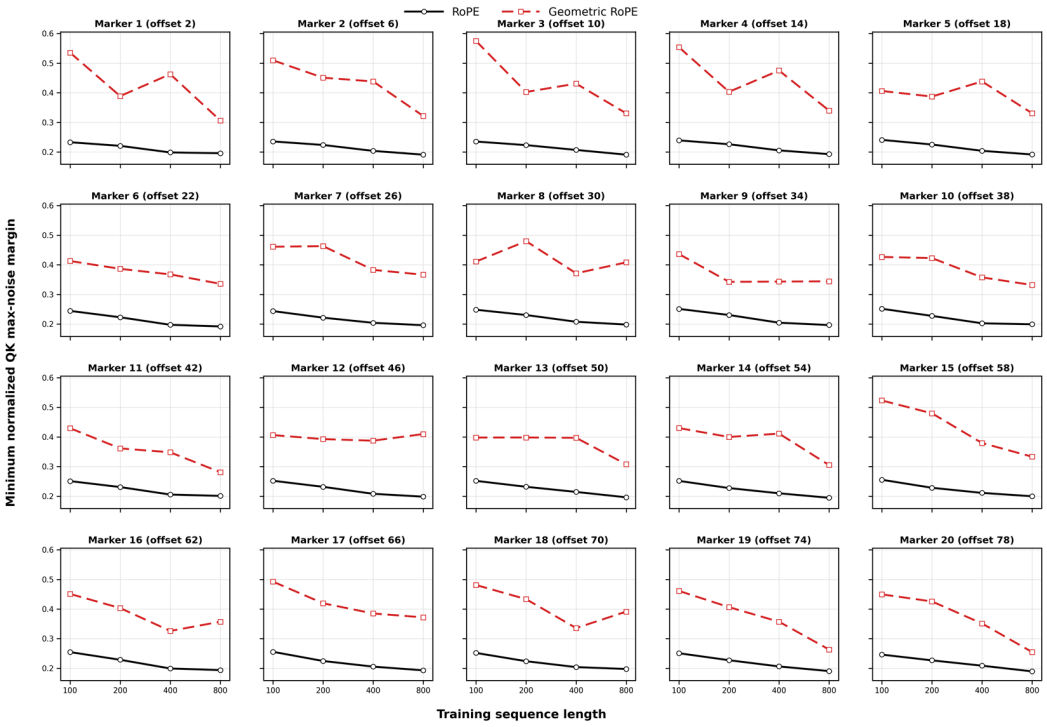

Figure 37: **Minimum normalized QK margins across training lengths.** The 4×5 grid shows markers 1–20, with each target offset given in its panel title. Horizontal axes are training lengths $L = 100, 200, 400, 800$; vertical axes show the reported minimum normalized maximum distractor margin. Black solid curves with circles denote standard RoPE, red dashed curves with squares denote Geometric RoPE.



Figure 38: **Difference in epochs to the low test loss.** The ordinate is the first entry epoch for Geometric RoPE minus that for standard RoPE, and the abscissa is training length.

by per layer validation loss, and computes position accuracy and MAE on all test tokens. Position curves average predictions per position over 100 shared test sequences for both models and plot 95% confidence intervals.