LARGE LANGUAGE MODELS WITHOUT QK-NORMALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Query-Key Normalization (QK-Norm) is widely adopted in modern large language models to control the growth of attention logits and improve training stability. However, query and key norms are not merely sources of instability. The query norm controls the sharpness of the attention distribution, while the key norm determines how strongly a token attracts different queries. These norms therefore carry token-specific information about attention selectivity and token importance. By normalizing both, QK-Norm suppresses this information. This raises a natural question: *can we control query-key representations without discarding their sensitivity to magnitude?* To address this question, we propose **QK-MAP**, a **M**agnitude-**A**ware **P**rojection for constructing queries and keys. Rather than removing magnitude through normalization, QK-MAP transforms the input representation using multiple learnable frequencies, so that changes in magnitude are encoded as bounded phase patterns. This allows magnitude to be represented through phase and direction rather than vector length alone. Its periodic features have a fixed norm while remaining sensitive to magnitude changes, separating magnitude encoding from norm growth. Multiple frequencies further help distinguish magnitudes and reduce periodic ambiguity. Experiments across model scales from 190M to 7B parameters show that QK-MAP achieves lower training and validation loss than both Standard attention and QK-Norm. Further analyses show that QK-MAP preserves sensitivity to input scale, assigns less attention to the first token, and produces stronger attention contrast for relevant evidence in synthetic long-context tests. These results demonstrate that query-key scale can be effectively controlled without sacrificing token-specific magnitude information. Our code is available at https://anonymous.4open.science/r/QK-MAP/.

1 INTRODUCTION

Query-Key Normalization (QK-Norm) has become a common component in modern large language models, including OLMo 3 and Qwen 3 (OLMo Team, 2025b; Yang et al., 2025). It addresses a common training problem: query and key norms can grow during training and produce increasingly large attention logits. Large logits can saturate the softmax, reduce attention entropy, and make optimization unstable (Dehghani et al., 2023; Zhai et al., 2023; Wortsman et al., 2023). QK-Norm controls this growth by normalizing projected queries and keys before computing their dot products (Henry et al., 2020; Rybakov et al., 2024). This prevents attention logits from growing directly with query and key norms and provides a practical way to improve training stability.

However, query and key norms do more than control the size of attention logits. They are also part of the attention function. For a query q_i and key k_j,

$$q_i^\top k_j = \|q_i\|_2 \|k_j\|_2 \cos\theta_{ij}. \tag{1}$$

The angle θ_{ij} describes how well the query and key directions match, while the two norms control the strength of this match. The query norm scales an entire row of attention logits. It therefore acts as a query-dependent temperature: a larger query norm generally produces a sharper attention distribution. The key norm scales a column of logits and changes one token's scores across different queries. The two norms thus play different roles and carry token-specific information within attention.

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

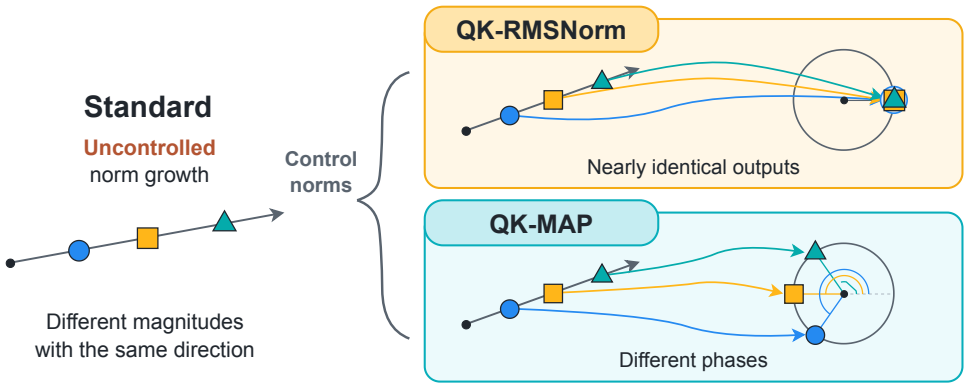

Figure 1: Schematic comparison for inputs with the same direction and different magnitudes. Standard attention preserves magnitude differences but allows query-key norms to grow. QK-RMSNorm controls norms but maps these inputs to nearly the same representation. QK-MAP's periodic features change with input magnitude while keeping their norm fixed.

Controlling these norms therefore involves a choice. Standard attention preserves this token-specific magnitude information, but allows query and key norms to grow. QK-Norm controls this growth, but makes the resulting representations much less sensitive to magnitude. This motivates the main question of this work: *Can we control query-key representations without losing their sensitivity to magnitude?*

To address this question, we propose **QK-MAP**, a **M**agnitude-**A**ware **P**rojection for constructing queries and keys. QK-MAP first maps an input representation into periodic features and then uses these features to construct queries and keys. When the input magnitude changes, the phases and coordinates of these features change while their overall norm remains fixed. Magnitude can therefore be represented by the pattern of the representation rather than only by its vector length. In this way, QK-MAP keeps query-key representations sensitive to magnitude without requiring their feature norms to grow with the input.

Figure 1 illustrates the main idea. Consider several inputs that have the same direction but different magnitudes. After QK-Norm, they produce nearly the same representation. With QK-MAP, changing the input magnitude changes the phase and coordinate pattern of the representation, even though its feature norm does not change. QK-MAP therefore moves magnitude information from vector length into the representation pattern.

We evaluate QK-MAP in language models ranging from 190M to 7B parameters under closely matched pretraining settings. Across model scales, QK-MAP achieves lower training and validation loss than both Standard attention and QK-Norm. At 7B parameters, it also obtains the best average accuracy on the main eight-task evaluation suite. Our analyses further show that QK-MAP retains sensitivity to input scale, assigns less attention to the first token, and gives relevant evidence stronger attention contrast in synthetic long-context tests. These results show that controlling query-key scale does not require removing token-specific magnitude information.

Our contributions are:

- We explain the different roles of query and key norms in attention and identify a tradeoff in QK-Norm: it controls logit growth but also removes much of the token-specific magnitude information carried by queries and keys.
- We introduce QK-MAP, which represents changes in input magnitude through fixed-norm phase and coordinate patterns. This keeps query-key representations sensitive to magnitude without relying only on growth in vector length.
- We evaluate QK-MAP from 190M to 7B parameters. The results show lower training and validation loss, favorable downstream performance, and clear changes in scale sensitivity, positional attention allocation, and relevant-evidence contrast.

2 MAGNITUDE INFORMATION IN ATTENTION

2.1 MAGNITUDE AS AN ATTENTION SIGNAL

For one attention head with width d_h, let

$$q_i = x_i W_Q, \qquad k_j = x_j W_K, \qquad S_{ij} = \frac{q_i^\top k_j}{\sqrt{d_h}}. \tag{2}$$

The attention distribution for query i is $a_i = \mathrm{softmax}(S_{i,:})$. Writing $q_i = r_i^q u_i^q$ and $k_j = r_j^k u_j^k$, where u_i^q and u_j^k have unit norm, gives

$$S_{ij} = \frac{r_i^q r_j^k}{\sqrt{d_h}} (u_i^q)^\top u_j^k. \tag{3}$$

The directional term measures how well a query and key match, while their magnitudes control how strongly this match is expressed. For fixed keys, r_i^q scales the entire i-th row of the logit matrix. It therefore acts as a query-dependent inverse temperature: increasing r_i^q makes the attention distribution sharper unless all logits are equal. For fixed queries, r_j^k scales the j-th column and changes token j's scores across different queries. Query and key magnitudes therefore carry different token-specific signals, although either can also produce overly large logits.

These two effects are complementary. Query magnitude controls how selectively a token reads from context, while key magnitude controls how strongly that token is considered by other queries. QK-Norm suppresses both forms of token-specific control.

2.2 NORM CONTROL VERSUS MAGNITUDE SENSITIVITY

A direct way to control query and key norms is ℓ_2 normalization:

$$\mathcal{L}(z) = \frac{z}{\|z\|_2}. \tag{4}$$

It fixes the output norm to one, preventing the input magnitude from directly increasing the attention logits. However, for any fixed nonzero direction u and $r > 0$,

$$\mathcal{L}(ru) = \mathcal{L}(u). \tag{5}$$

ℓ_2 normalization therefore controls norm growth by making all positive rescalings along the same direction indistinguishable.

Modern language models commonly use RMSNorm instead of exact ℓ_2 normalization:

$$\mathcal{N}_\epsilon(z) = \frac{z}{\sqrt{\|z\|_2^2/d_h + \epsilon}} \odot g, \tag{6}$$

where g is a learned gain shared across tokens (Zhang & Sennrich, 2019). When $\epsilon = 0$ and the gain is ignored, $\mathcal{N}_0(z)$ differs from ℓ_2 normalization only by the constant factor $\sqrt{d_h}$. It therefore has the same invariance to input magnitude.

Proposition 1 (Magnitude information after QK-RMSNorm)**.** *Let $z = ru$, where $u \neq 0$ is fixed and $r > 0$. For $\epsilon = 0$,*

$$\mathcal{N}_0(ru) = \mathcal{N}_0(u). \tag{7}$$

For $\epsilon > 0$,

$$\left\| \frac{\partial \mathcal{N}_\epsilon(ru)}{\partial r} \right\|_2 = \frac{\epsilon \|u\|_2}{(r^2 c + \epsilon)^{3/2}}, \qquad c = \frac{\|u\|_2^2}{d_h}, \tag{8}$$

where the fixed gain g is omitted for clarity. The dependence on magnitude therefore decreases rapidly as r grows.

QK-L2Norm fixes the norm exactly, while QK-RMSNorm allows learned dimension-wise scaling through g and uses a finite ϵ for numerical stability. Despite this difference, both prevent token magnitude from directly controlling the query or key norm. The gain g is shared across tokens, so

it cannot restore a different magnitude signal for each query or key. Thus, the same mechanism that controls token-wise norm growth also removes most of the information carried by changes in magnitude.

QUEST (Govindarajan et al., 2026) takes an asymmetric approach: it normalizes keys to unit length but leaves queries unnormalized. It therefore retains query-dependent attention sharpness while fixing the key-norm factor in Equation 3. We include QUEST in the key-side comparison in Figure 2.

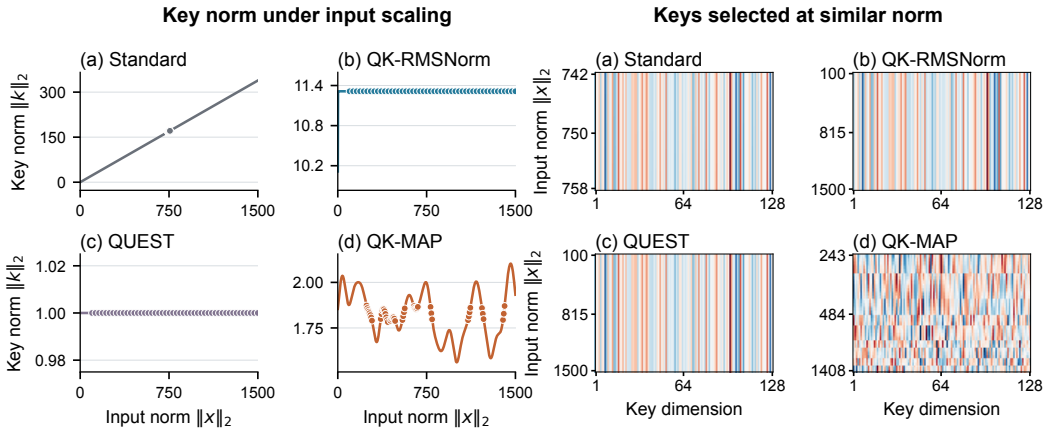

Figure 2: Key-side response to input scaling. *Left:* Key norms as input norm varies; dots mark samples within a narrow, method-specific key-norm range. *Right:* Key coordinates of the selected samples, ordered by input norm. Each heatmap uses its own color scale.

3 QK-MAP

QK-MAP replaces QK normalization with a learned periodic representation. Its goal is to keep the query-key path responsive to input magnitude without requiring that information to be carried only by vector length.

Let $x \in \mathbb{R}^D$ be the representation entering the attention block. QK-MAP learns $W_p \in \mathbb{R}^{D\times m}$ and $b_p \in \mathbb{R}^m$, and computes

$$\Phi(x) = \begin{bmatrix} \cos(xW_p + b_p) \\ \sin(xW_p + b_p) \end{bmatrix} \in \mathbb{R}^{2m}. \tag{9}$$

The sine and cosine are applied elementwise. In our main configuration, $m = D/2$, so the output dimension of $\Phi(x)$ is also D.

The same periodic representation is used to form queries and keys, with separate projections for each path:

$$q = \Phi(x)W_Q, \qquad k = \Phi(x)W_K, \qquad v = xW_V. \tag{10}$$

The value path is unchanged. After projection, queries and keys are reshaped into heads, and RoPE is applied as usual. The frequencies in W_p are learned jointly with the language model. Figure 3 compares where the two methods control query-key scale.

The design is inspired by the representation idea behind RoPE (Su et al., 2024). RoPE represents position through rotations at multiple base frequencies: each coordinate remains bounded, while different positions produce different combinations of phases over the context range. QK-MAP applies a similar idea to magnitude. Instead of using fixed frequencies to encode a position index, it uses learned frequencies to turn changes in the input representation into changes across multiple phase coordinates. In this way, a potentially unbounded magnitude can be represented by a bounded pattern. This use differs from random and learned Fourier features for kernel approximation or coordinate encoding (Rahimi & Recht, 2007; Tancik et al., 2020; Li et al., 2021). Here the periodic map is learned inside each attention block and is applied to the token representation rather than to a position index.

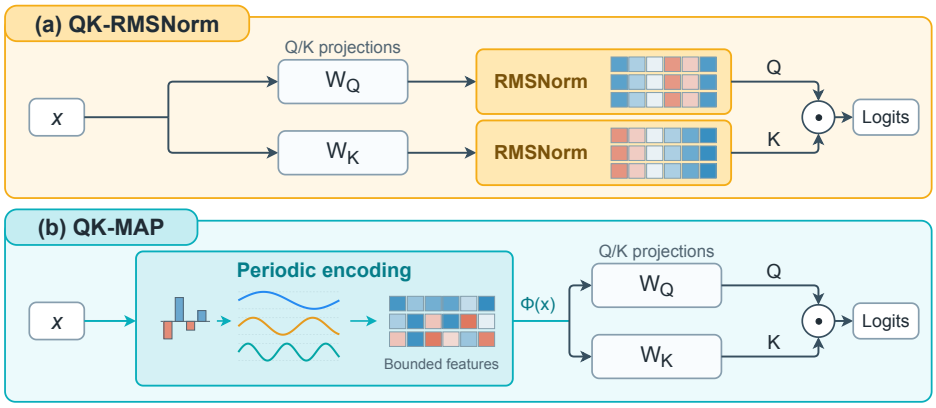

Figure 3: Comparison of QK-RMSNorm and QK-MAP. QK-RMSNorm normalizes projected queries and keys before attention. QK-MAP first forms a shared periodic representation and then applies separate query and key projections.

The key property is stated below.

Proposition 2 (Magnitude encoding without norm growth)**.** *For the feature map in Equation 9,*

$$\|\Phi(x)\|_2^2 = m \qquad \textit{for every } x. \tag{11}$$

For a fixed direction u and $x = ru$,

$$\left\| \frac{\partial \Phi(ru)}{\partial r} \right\|_2^2 = \|uW_p\|_2^2. \tag{12}$$

Therefore, the feature norm remains fixed, while the feature values continue to change with input magnitude whenever $uW_p \neq 0$.

This result gives the main difference from QK-Norm. QK-Norm makes representations with the same direction nearly independent of their magnitude. QK-MAP instead moves magnitude information from vector length into the coordinates of the representation.

The use of multiple frequencies also matters. A single phase coordinate repeats as the input magnitude changes. With multiple frequencies, the complete feature vector repeats only when all phase coordinates repeat at the same time. For a fixed direction u, two magnitudes $r_1 \neq r_2$ produce the same periodic features only if

$$(r_1 - r_2)u^\top w_j \in 2\pi\mathbb{Z} \qquad \text{for every } j. \tag{13}$$

Each phase coordinate is periodic, but multiple learned frequencies provide a richer collection of phase patterns through which the model can represent magnitude.

Figure 2 illustrates these properties on the key side so that QUEST, which normalizes only keys, can be compared directly. For a fixed direction u, Standard key norms grow with the input magnitude r. QK-RMSNorm and QUEST keep the key norm nearly constant, whereas QK-MAP keeps it bounded while still responding to r. The right block compares inputs whose key norms are similar within each method. The normalized baselines give nearly identical key coordinates across widely separated input magnitudes. QK-MAP instead gives different coordinate patterns even when its key norms are similar. Thus, controlling key norm need not remove all information about input magnitude.

4 Experiments

4.1 Setup

We compare Standard attention, QK-Norm, and QK-MAP in decoder-only language models from 190M to 7B parameters. Within each scale, all variants use the same backbone, pretraining data,

token budget, global batch size, and initialization and data seeds. The pretraining data follow the Dolma-based OLMo mixture (Groeneveld et al., 2024; Soldaini et al., 2024). The cross-scale comparison uses a learned per-dimension residual variant of QK-MAP, with a slightly reduced feed-forward width to match parameter counts. Architecture and residual ablations are in Appendix E.

For downstream evaluation, we follow the OLMES protocol (Gu et al., 2025b) and report Avg-8, the unweighted mean accuracy over ARC-Easy (Clark et al., 2018), PIQA (Bisk et al., 2020), SciQ (Welbl et al., 2017), ARC-Challenge (Clark et al., 2018), COPA (Roemmele et al., 2011), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), and WinoGrande (Sakaguchi et al., 2020). For language modeling, we report final-window training cross-entropy and validation cross-entropy on C4 (Raffel et al., 2020), Pile (Gao et al., 2020), M2D2-S2ORC (Reid et al., 2022), and six Dolma domains (Soldaini et al., 2024). We then examine magnitude use, attention allocation, and performance under text perturbations. Full configurations and evaluation protocols are in Appendices C and D.

4.2 Main Results

Scale	Variant	CE↓		Downstream evaluation (%↑)								
		Train	Val.	ARC-E	PIQA	SciQ	ARC-C	COPA	HS	OBQA	WG	Avg-8
190M	Standard	3.86	4.19	27.72	58.10	30.30	23.75	56.00	27.40	27.80	50.90	37.75
	QK-Norm	3.78	4.10	**35.79**	**60.70**	**34.10**	**25.08**	58.00	27.60	**30.00**	48.00	**39.91**
	QK-MAP	**3.67**	**3.98**	30.74	60.62	32.93	23.22	**61.00**	**28.20**	26.33	**51.03**	39.26
370M	Standard	3.19	3.54	39.12	58.90	67.00	**31.10**	60.00	25.90	28.80	**52.00**	45.35
	QK-Norm	3.17	3.52	**40.70**	**59.40**	**70.00**	23.75	**64.00**	27.00	33.60	48.80	45.91
	QK-MAP	**3.10**	**3.45**	**40.70**	58.00	69.70	25.42	61.00	**27.30**	**35.60**	49.80	**45.94**
760M	Standard	2.79	3.16	**47.72**	62.00	79.80	**31.44**	64.00	32.30	34.40	**50.30**	**50.24**
	QK-Norm	2.79	3.16	47.19	61.70	77.50	29.77	**65.00**	32.30	33.60	49.80	49.61
	QK-MAP	**2.76**	**3.13**	45.09	**62.50**	**79.90**	28.43	64.00	**33.20**	**34.80**	49.90	49.73
1B	Standard	2.69	3.09	48.42	**63.90**	**82.00**	27.76	66.00	35.00	36.00	**51.70**	51.35
	QK-Norm	2.70	3.09	**51.05**	63.40	79.40	31.77	66.00	**35.70**	**37.40**	51.50	52.03
	QK-MAP	**2.67**	**3.07**	49.30	**63.90**	**82.00**	**34.78**	**67.00**	34.20	34.40	51.30	**52.11**
7B	Standard	2.25	2.89	49.65	**68.00**	**84.80**	37.12	69.00	46.60	39.00	**52.70**	55.86
	QK-Norm	2.24	2.90	**54.21**	67.90	84.20	32.44	**72.00**	45.30	37.80	50.00	55.48
	QK-MAP	**2.23**	**2.86**	52.81	**68.00**	83.00	**39.13**	70.00	**47.50**	**41.20**	52.40	**56.75**

Table 1: Language-modeling cross-entropy and downstream accuracy from 190M to 7B.

Language modeling and downstream performance. Table 1 shows that QK-MAP achieves the lowest training and validation CE at every evaluated scale. The validation improvement holds across C4, Pile, M2D2-S2ORC, and the Dolma domains. QK-MAP also improves average downstream accuracy at most model scales, extending its benefits beyond language modeling loss.

Training dynamics. Figure 4 shows the training and C4 validation trajectories at 7B. The late-training views make the persistent CE advantage of QK-MAP visible beyond the final checkpoint. Panel (c) shows that the downstream result is also consistent across nearby checkpoints: QK-MAP ranks first in Avg-8 at all 16 common checkpoints evaluated near the end of training (Appendix B.2).

4.3 Do the Models Use Magnitude Information?

If key magnitudes aid prediction, assigning them to different tokens should increase loss. We shuffle key norms across sequences at the same position before attention (Figure 5(a)), keeping directions and the set of norms fixed. We compare CE before and after shuffling on 128 C4 sequences, using 190M checkpoints and the direct, same-dimensional QK-MAP variant. Appendix G.1 gives settings and confidence intervals.

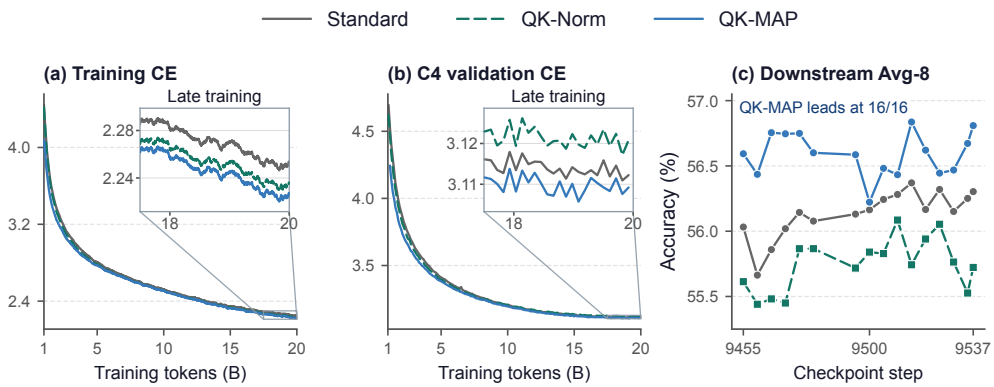

Figure 4: Training dynamics of the 7B models. (a) Training CE with 100-update smoothing. (b) C4 validation CE, evaluated every 50 updates. Insets and connecting lines show the final stage of training. (c) Avg-8 at 16 common checkpoints near the end of training.

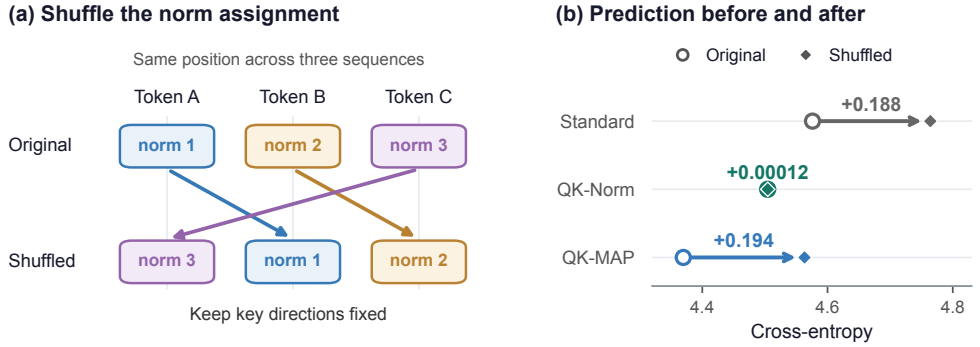

Figure 5: Key-norm intervention at 190M. (a) Reassigning norms among tokens; norm values are illustrative. (b) CE before and after shuffling, with the increase labeled.

Shuffling key norms raises CE by 0.194 for QK-MAP and 0.188 for Standard, but only 0.00012 for QK-Norm (Figure 5(b)). The loss increase shows that QK-MAP uses token-specific key magnitudes for prediction, as Standard attention does.

4.4 Where Does Attention Go?

We next examine how QK-MAP allocates attention in the 7B models. Two complementary measurements track attention to the first key and attention to inserted evidence.

Less concentration on the first key. Initial tokens often receive large attention weights, a behavior associated with attention sinks (Xiao et al., 2024; Gu et al., 2025a). We measure the percentage of attention assigned to the first key, averaged over layers, heads, and query positions. Figure 6(a) gives a direct view: the strong vertical attention bands in the PeS2o baseline maps are weaker under QK-MAP. Panel (b) summarizes the relative reduction in mean first-key attention over eleven domains at context lengths from 32 to 512. QK-MAP has the lowest first-key attention at every length and is below both baselines in all 55 domain–length comparisons. Its mean is 9.1%–12.7% lower than Standard. Domain-level results and attention maps are in Appendix F.2.

Stronger attention to relevant evidence. To examine content allocation, we insert evidence into distractor text and measure evidence-to-distractor attention SNR for the final question-token query. We average the statistic across all layers. Such synthetic prompts offer a controlled setting for studying attention to relevant content (Liu et al., 2024; Hsieh et al., 2024; Bai et al., 2023; Modarressi

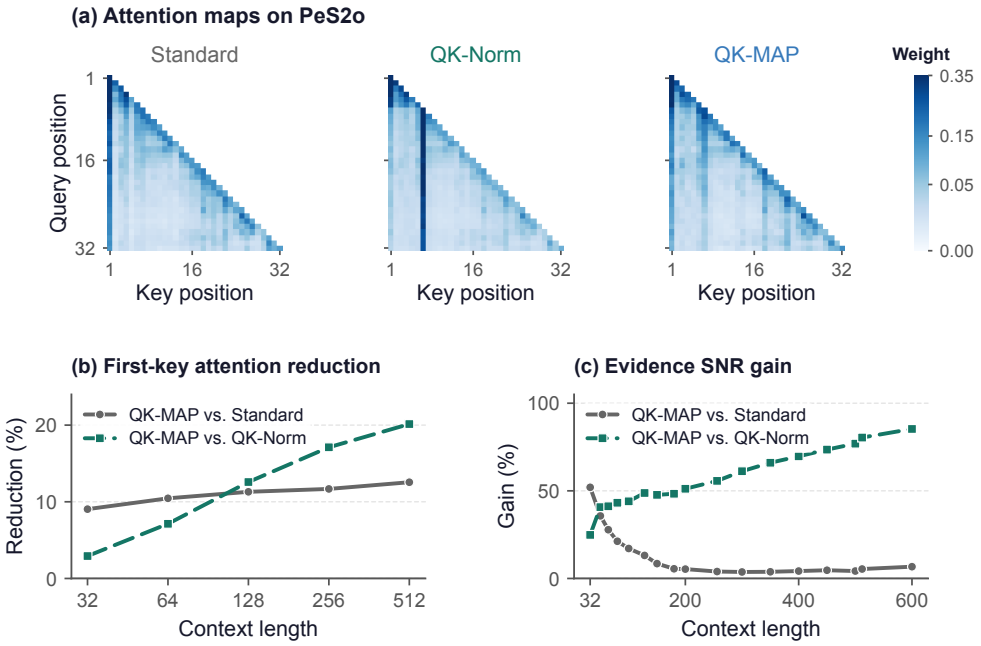

Figure 6: Attention allocation in the 7B models. (a) PeS2o attention maps over the first 32 positions, with a shared square-root color scale. (b) Relative reduction in mean first-key attention over eleven domains. (c) Relative gain in average-layer evidence SNR. Panels (b)–(c) show QK-MAP relative to each baseline.

et al., 2025). Figure 6(c) shows that QK-MAP has the highest average-layer SNR at every evaluated context length. Its relative gain over QK-Norm grows from 24.9% at 32 tokens to 85.3% at 600 tokens. The pattern also holds when the prompt contains more evidence or different task types, as shown in Table 2.

Prompt set	Standard	QK-Norm	QK-MAP
1 needle	0.272	0.238	**0.338**
2 needles	0.352	0.318	**0.426**
3 needles	0.418	0.415	**0.506**
Multi-task	0.342	0.290	**0.386**

Table 2: Average-layer evidence attention SNR at 7B. Higher values indicate stronger contrast between evidence and distractors.

Together, these results show two consistent changes: QK-MAP assigns less attention to the first key and gives relevant evidence stronger attention contrast. The evidence result extends from single-needle prompts to multiple needles and mixed tasks, rather than depending on one prompt format. Appendix F.1 gives the prompt settings and additional layer-pooling results.

4.5 Performance under Text Perturbations

We also test whether the language-modeling advantage persists when the input text is perturbed. Following the word-replacement test used by QUEST (Govindarajan et al., 2026), we randomly replace words with `AAA` in a fixed WikiText-103 validation sample (Merity et al., 2016). We evaluate Standard, QK-Norm, QUEST, and the direct QK-MAP variant at 190M using identical text and replacements. The replacement rates are 1.5%, 2.5%, and 5%; Appendix G.2 gives the evaluation protocol.

Variant	Clean	1.5%	2.5%	5%
Standard	122.77	141.10	155.51	191.67
QK-Norm	111.77	129.56	143.63	177.74
QUEST	116.07	134.38	148.94	184.26
QK-MAP	**95.60**	**112.12**	**124.12**	**153.27**

Table 3: Perplexity on clean and word-replaced WikiText-103 at 190M.

QK-MAP achieves the lowest perplexity on clean text and at every tested replacement rate (Table 3). At 5% replacement, its perplexity is 13.8% below QK-Norm and 16.8% below QUEST. Its language-modeling advantage therefore persists as local word information is disrupted.

5 Related Work

Attention Scaling and Normalization. Scaled dot-product attention divides logits by $\sqrt{d_h}$, which controls their scale at initialization but does not bound later growth (Vaswani et al., 2017). Transformer stability has also been studied through normalization placement and residual scaling (Xiong et al., 2020). ScaleNorm, NormFormer, and DeepNorm provide further normalization- and residual-based designs (Nguyen & Salazar, 2019; Shleifer et al., 2021; Wang et al., 2022). Other studies trace training failures to loss spikes, growing attention logits, or entropy collapse (Takase et al., 2023; Dehghani et al., 2023; Zhai et al., 2023; Wortsman et al., 2023). These methods mainly stabilize the wider Transformer block. A more direct line controls the attention scores themselves. Examples include Lipschitz-normalized attention (Kim et al., 2021; Dasoulas et al., 2021), scaled cosine attention (Henry et al., 2020; Liu et al., 2022), logit soft capping (Gemma Team et al., 2024; Rybakov et al., 2024), and methods that limit changes in attention logits during optimization (Anson & Aitchison, 2025). QK-Norm is now used in several large-model architectures (Chameleon Team, 2024; OLMo Team, 2025a;b; Yang et al., 2025; Gemma Team et al., 2025), and has also been adapted to latent-attention designs (Han et al., 2026).

These controls improve stability, but they also change which input properties can affect attention. QKNorm replaces raw dot products with scaled cosine similarity (Henry et al., 2020), so radial changes in a projected query or key are removed. Recent work has started to preserve selected magnitude effects. QUEST keeps a query-dependent temperature while normalizing keys (Govindarajan et al., 2026), while Norm$\times$Direction restores query magnitude in linear attention (Meng et al., 2025). QK-MAP addresses the same broad trade-off from a different direction: it keeps both query- and key-side input changes available through a shared bounded representation, then learns separate query and key projections.

Periodic representations in neural networks. Periodic features have been used for several different goals. Random Fourier features approximate shift-invariant kernels (Rahimi & Recht, 2007), and Performer uses positive random features to approximate softmax attention (Choromanski et al., 2021). Fourier input mappings and periodic activations improve the representation of high-frequency signals (Tancik et al., 2020; Sitzmann et al., 2020). Learned Fourier features encode spatial coordinates (Li et al., 2021), while RoPE uses paired rotations to encode relative position in attention (Su et al., 2024). Time2Vec similarly maps a scalar into learned linear and periodic coordinates (Kazemi et al., 2019). FAN further combines learned sine–cosine features with non-periodic features as a general network component (Dong et al., 2025). QK-MAP uses a related paired construction, but its purpose is different: it does not approximate an attention kernel or encode token position. Instead, it learns a bounded multi-frequency representation of the input to the query and key projections.

6 Conclusion

QK-Norm controls attention-logit growth, but also removes much of the token-specific magnitude information carried by queries and keys. We introduced QK-MAP, which uses learned multi-frequency periodic features to represent magnitude without relying on feature-norm growth. Experiments from 190M to 7B show consistent improvements in language modeling. Attention analyses further show lower first-key attention and stronger attention contrast for relevant evidence. Overall, our results show that query-key scale can be controlled without discarding magnitude information, making QK-MAP a practical alternative to QK normalization.

AI USE STATEMENT

In preparing this manuscript, we use LLMs to aid and polish the writing. Specifically, LLMs improve clarity, grammar, and phrasing, ensuring the text is concise and readable. The use of LLMs does not influence the technical contributions or the interpretation of experimental findings. All content polished by LLMs is carefully checked by the authors.

REPRODUCIBILITY STATEMENT

We provide anonymous source code for the QK-MAP module, OLMo integration helpers, training-configuration scaffolding, analysis probes, and evaluation scripts at https://anonymous.4open.science/r/QK-MAP/. Matched training settings, residual options, and compute notes are given in Appendix C. Downstream metrics, validation corpora, and text-perturbation protocols are specified in Appendices D and G. Proofs of radial sensitivity and bounded logits appear in Appendix A, and architecture and residual ablations in Appendix E. Within each scale, all variants share the same backbone, data, token budget, batch size, and seeds, as described in Section 4.

REFERENCES

Ben Anson and Laurence Aitchison. Controlling changes to attention logits. *arXiv preprint arXiv:2511.21377*, 2025.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2020.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

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. In *arXiv preprint arXiv:1803.05457*, 2018.

George Dasoulas, Kevin Scaman, and Aladin Virmaux. Lipschitz normalization for self-attention layers with application to graph neural networks. In *International Conference on Machine Learning*, pp. 2456–2466, 2021.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, et al. Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv:2302.05442*, 2023.

Yihong Dong, Ge Li, Yongding Tao, Xue Jiang, Kechi Zhang, Jia Li, Jinliang Deng, Jing Su, Jun Zhang, and Jingjing Xu. FAN: Fourier analysis networks. In *Advances in Neural Information Processing Systems*, 2025.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gemma Team, Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. QUEST: A robust attention formulation using query-modulated spherical attention. In *International Conference on Learning Representations*, 2026.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15789–15809, 2024. doi: 10.18653/v1/2024.acl-long.841.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, 2025a.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5020–5048, 2025b. doi: 10.18653/v1/2025.findings-naacl.282.

Yizhou Han, Yao Zhao, Jun Zhou, Longfei Li, and Ruoyu Sun. QK-normed MLA: QK normalization without full key caching. *arXiv preprint arXiv:2606.16310*, 2026.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020. doi: 10.18653/v1/2020.findings-emnlp.379.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571, 2021.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022.

Weikang Meng, Yadan Luo, Liangyu Huo, Yingjian Li, Yaowei Wang, Xin Li, and Zheng Zhang. Norm×direction: Restoring the missing query norm in vision linear attention. *arXiv preprint arXiv:2506.21137*, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of EMNLP*, 2018.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*, 2025.

Toan Q. Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of self-attention. *arXiv preprint arXiv:1910.05895*, 2019.

OLMo Team. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025a.

OLMo Team. OLMo 3. *arXiv preprint arXiv:2512.13961*, 2025b.

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.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. M2D2: A massively multi-domain language modeling dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 964–975, 2022. doi: 10.18653/v1/2022.emnlp-main.63.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium*, 2011.

Oleg Rybakov, Mike Chrzanowski, Peter Dykas, Jinze Xue, and Ben Lanir. Methods of improving LLM training stability. *arXiv preprint arXiv:2410.16682*, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *AAAI Conference on Artificial Intelligence*, 2020.

Sam Shleifer, Jason Weston, and Myle Ott. Normformer: Improved transformer pretraining with extra normalization. *arXiv preprint arXiv:2110.09456*, 2021.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15725–15788, 2024.

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

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. Spike no more: Stabilizing the pre-training of large language models. *arXiv preprint arXiv:2312.16903*, 2023.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. *arXiv preprint arXiv:2203.00555*, 2022.

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.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pp. 10524–10533, 2020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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*, 2019.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

APPENDIX CONTENTS

A Theoretical Details **15**

A.1 Setup and Assumptions 15

A.2 Radial Information Lost by QK-RMSNorm 15

A.3 Bounded and Magnitude-Sensitive QK-MAP 16

A.4 When Different Magnitudes Produce the Same Features 17

B Experimental Setup **18**

B.1 Downstream Evaluation 18

B.2 Nearby-Checkpoint Evaluation 18

C Detailed Configurations and Reproducibility **18**

C.1 Scale-Specific Configurations and Parameter Matching 18

C.2 Shared Optimization and Implementation 19

C.3 Released Implementation and Compute 19

D Evaluation Protocol **19**

E Architecture and Residual Ablations **20**

F Attention Diagnostics **20**

F.1 Synthetic Evidence Contrast 20

F.2 Cross-Domain First-Token Allocation 21

G Magnitude Intervention and Text Perturbation **23**

G.1 Norm-Shuffling Protocol 23

G.2 Word-Replacement Protocol 24

A Theoretical Details

QK-MAP has two design goals: preserving information about token magnitude and keeping attention logits bounded as that magnitude grows. We analyze both properties below.

A.1 Setup and Assumptions

We use column vectors. For a query token at position i, write $x_i = r_i u_i$, where $r_i > 0$ is its magnitude and $u_i \neq 0$ is a fixed direction. A key token is written in the same way as $y_j = t_j v_j$. The query and key projections are bias-free, as in our implementation. The RoPE matrices R_i and R_j are orthogonal, so they preserve Euclidean norms. The scaled attention logit is

$$s_{ij} = \frac{q_i^\top k_j}{\sqrt{d_h}}. \tag{14}$$

For QK-RMSNorm, let

$$\mathcal{N}_{\epsilon,G}(z) = \frac{Gz}{\sqrt{\|z\|_2^2/d_h + \epsilon}}, \tag{15}$$

where G is the diagonal learned gain and $\epsilon \geq 0$. The gain is shared across tokens and has finite operator norm.

A.2 Radial Information Lost by QK-RMSNorm

Proposition A.1 (Radial invariance and score sensitivity)**.** *Let $a = W_Q u_i \neq 0$, $c_a = \|a\|_2^2/d_h$, and*

$$q_i(r_i) = R_i \mathcal{N}_{\epsilon,G_Q}(r_i a).$$

For idealized normalization ($\epsilon = 0$), both $q_i(r_i)$ and s_{ij} are invariant to every positive rescaling of r_i. For $\epsilon > 0$,

$$\left\| \frac{\partial q_i}{\partial r_i} \right\|_2 \leq \frac{\epsilon \|G_Q\|_{\mathrm{op}} \|a\|_2}{(r_i^2 c_a + \epsilon)^{3/2}}, \tag{16}$$

$$\left| \frac{\partial s_{ij}}{\partial r_i} \right| \leq \frac{\epsilon \|G_Q\|_{\mathrm{op}} \|G_K\|_{\mathrm{op}} \|a\|_2}{(r_i^2 c_a + \epsilon)^{3/2}}. \tag{17}$$

An analogous bound holds for the key magnitude t_j. Once the projected query norm is outside the small-ϵ regime, the radial sensitivity of QK-RMSNorm decays as $O(r_i^{-3})$.

Proof. Substituting $W_Q x_i = r_i a$ into Equation 15 gives

$$q_i(r_i) = R_i G_Q \frac{r_i a}{\sqrt{r_i^2 c_a + \epsilon}}. \tag{18}$$

When $\epsilon = 0$ and $r_i > 0$, the scale cancels:

$$q_i(r_i) = R_i G_Q \frac{a}{\sqrt{c_a}}. \tag{19}$$

The query is therefore identical for all positive r_i. Holding the key fixed in Equation 14 makes the score invariant as well.

For $\epsilon > 0$, differentiating the scalar factor in Equation 18 yields

$$\frac{\partial q_i}{\partial r_i} = R_i G_Q a \left[\frac{1}{(r_i^2 c_a + \epsilon)^{1/2}} - \frac{r_i^2 c_a}{(r_i^2 c_a + \epsilon)^{3/2}} \right] \tag{20}$$

$$= R_i G_Q a \, \frac{\epsilon}{(r_i^2 c_a + \epsilon)^{3/2}}. \tag{21}$$

Because R_i is orthogonal, $\|R_i G_Q a\|_2 \leq \|G_Q\|_{\mathrm{op}} \|a\|_2$, which proves Equation 16.

It remains to transfer this result to the attention score. For any z,

$$\|\mathcal{N}_{\epsilon,G_K}(z)\|_2 \le \|G_K\|_{\mathrm{op}} \frac{\|z\|_2}{\sqrt{\|z\|_2^2/d_h + \epsilon}} \tag{22}$$

$$\le \sqrt{d_h}\|G_K\|_{\mathrm{op}}. \tag{23}$$

RoPE preserves this bound, so $\|k_j\|_2 \le \sqrt{d_h}\|G_K\|_{\mathrm{op}}$. The Cauchy–Schwarz inequality then gives

$$\left|\frac{\partial s_{ij}}{\partial r_i}\right| = \frac{1}{\sqrt{d_h}} \left| \left(\frac{\partial q_i}{\partial r_i}\right)^\top k_j \right| \tag{24}$$

$$\le \frac{1}{\sqrt{d_h}} \left\| \frac{\partial q_i}{\partial r_i} \right\|_2 \|k_j\|_2, \tag{25}$$

which, together with Equation 16, proves Equation 17. Interchanging query and key proves the corresponding statement for t_j. □

The proposition concerns the attention score itself, rather than only an intermediate normalized vector. A learned gain changes the constant in the bound, but cannot remove the exact invariance at $\epsilon = 0$ or the cubic decay for finite ϵ.

A.3 Bounded and Magnitude-Sensitive QK-MAP

Let $W_p = [w_1, \dots, w_m] \in \mathbb{R}^{D\times m}$ and $b \in \mathbb{R}^m$. QK-MAP uses

$$\theta(x) = W_p^\top x + b, \qquad \Phi(x) = \begin{bmatrix}\cos\theta(x) \\ \sin\theta(x)\end{bmatrix} \in \mathbb{R}^{2m}. \tag{26}$$

Proposition A.2 (Bounded features and directional sensitivity)**.** *For every $x \in \mathbb{R}^D$,*

$$\|\Phi(x)\|_2^2 = m. \tag{27}$$

Its Jacobian satisfies

$$J_\Phi(x)^\top J_\Phi(x) = W_p W_p^\top. \tag{28}$$

Consequently, for any perturbation δ,

$$\|J_\Phi(x)\delta\|_2^2 = \|W_p^\top \delta\|_2^2. \tag{29}$$

In particular, along $x = ru$,

$$\left\| \frac{\partial \Phi(ru)}{\partial r} \right\|_2 = \|W_p^\top u\|_2, \tag{30}$$

which is positive whenever $u \notin \ker(W_p^\top)$ and does not decay as r grows.

Proof. Writing $\theta_j = w_j^\top x + b_j$, each paired coordinate obeys $\cos^2\theta_j + \sin^2\theta_j = 1$. Summing the m pairs proves Equation 27. Elementwise differentiation gives

$$J_\Phi(x) = \begin{bmatrix} -\operatorname{Diag}(\sin\theta) W_p^\top \\ \operatorname{Diag}(\cos\theta) W_p^\top \end{bmatrix}. \tag{31}$$

Therefore,

$$J_\Phi(x)^\top J_\Phi(x) = W_p \operatorname{Diag}(\sin^2\theta) W_p^\top \tag{32}$$

$$+ W_p \operatorname{Diag}(\cos^2\theta) W_p^\top \tag{33}$$

$$= W_p W_p^\top, \tag{34}$$

which proves Equation 28. Multiplying on the left and right by $\delta^\top$ and δ proves Equation 29. Setting $\delta = u$ gives the radial result. □

Proposition A.3 (Magnitude-independent logit bound)**.** *Let*

$$q_i = R_i A_Q \Phi(x_i), \qquad k_j = R_j A_K \Phi(y_j),$$

where A_Q and A_K are learned linear projections. Then

$$|s_{ij}| \le \frac{m}{\sqrt{d_h}} \|A_Q\|_{\mathrm{op}} \|A_K\|_{\mathrm{op}}, \tag{35}$$

independently of $\|x_i\|_2$ and $\|y_j\|_2$. The mapped query remains radially sensitive whenever $A_Q J_\Phi(x_i) u_i \neq 0$, with the analogous condition for the key.

Proof. Equation 27 and orthogonality of RoPE give

$$\|q_i\|_2 \le \sqrt{m} \|A_Q\|_{\mathrm{op}}, \qquad \|k_j\|_2 \le \sqrt{m} \|A_K\|_{\mathrm{op}}.$$

Applying Cauchy–Schwarz to Equation 14 proves Equation 35. The chain rule gives $\partial q_i / \partial r_i = R_i A_Q J_\Phi(x_i) u_i$, which is nonzero under the stated condition. □

The default implementation adds a learned residual,

$$\Psi_\alpha(x) = \Phi(x) + \mathrm{Diag}(\alpha) x, \tag{36}$$

with $2m = D$ and α initialized at zero. The attention-block input is produced by a preceding RMSNorm. If $\|x\|_2 \le B$, then

$$\|\Psi_\alpha(x)\|_2 \le \sqrt{m} + \|\alpha\|_\infty B. \tag{37}$$

Thus, the residual variant starts exactly as the bounded periodic map and remains bounded on the normalized input domain. This connects the theoretical map to the residual design used in the scale experiments.

A.4 When Different Magnitudes Produce the Same Features

For a fixed direction u, define radial frequencies $\omega_j = w_j^\top u$. Equality of the paired sine and cosine features is equivalent to

$$(r_1 - r_2)\omega_j \in 2\pi\mathbb{Z} \qquad \text{for every } j. \tag{38}$$

Proposition A.4 (A sufficient condition for global injectivity)**.** *Suppose two nonzero radial frequencies satisfy $\omega_p/\omega_q \notin \mathbb{Q}$. Then the map $r \mapsto \Phi(ru)$ is injective on $\mathbb{R}$.*

Proof. Assume that $r_1 \neq r_2$ produces a collision and let $\Delta = r_1 - r_2 \neq 0$. Equation 38 implies $\Delta\omega_p = 2\pi n_p$ and $\Delta\omega_q = 2\pi n_q$ for integers n_p, n_q. Since Δ, ω_p, and ω_q are nonzero, both integers are nonzero. Dividing the two equalities gives $\omega_p/\omega_q = n_p/n_q \in \mathbb{Q}$, contradicting the assumption. □

If W_p is drawn from a jointly absolutely continuous distribution, the rational-ratio failure set has measure zero for any fixed nonzero u and $m \ge 2$. Hence, the proposition holds almost surely at initialization. Learning may change the frequencies, so this is a sufficient condition rather than an unconditional guarantee.

Proposition A.5 (Separation on a bounded magnitude range)**.** *Assume the condition of Proposition A.4. For any compact interval $I \subset \mathbb{R}$ and any $\eta > 0$, there exists $\gamma(I, \eta) > 0$ such that*

$$\|\Phi(r_1 u) - \Phi(r_2 u)\|_2 \ge \gamma(I, \eta) \tag{39}$$

for all $r_1, r_2 \in I$ with $|r_1 - r_2| \ge \eta$.

Proof. The set $K = \{(r_1, r_2) \in I^2 : |r_1 - r_2| \ge \eta\}$ is compact. If K is empty, the claim holds for any $\gamma > 0$. Otherwise, the function $f(r_1, r_2) = \|\Phi(r_1 u) - \Phi(r_2 u)\|_2$ is continuous. By injectivity, f is strictly positive on K. The extreme-value theorem therefore gives a positive attained minimum $\gamma(I, \eta) = \min_K f$. □

This final result states the practical role of multiple frequencies: on any finite magnitude range, inputs separated by a fixed amount also have a strictly positive feature separation.

Scale	L	D	H	d_h	$\mathrm{FF}_{\mathrm{base}}$	$\mathrm{FF}_{\mathrm{MAP}}$	Tokens	Steps	Peak LR
190M	12	768	12	64	3072	2944	2B	954	6×10^{-4}
370M	16	1024	16	64	4096	3928	4B	1908	6×10^{-4}
760M	16	1536	16	96	6144	5888	8B	3815	6×10^{-4}
1B	18	2048	16	128	8192	7852	10B	4769	4×10^{-4}
7B	32	4096	32	128	11008	10328	20B	9537	3×10^{-4}

Table 4: Backbone and schedule by scale. L is the number of layers, D the hidden size, H the number of heads, and d_h the head width. $\mathrm{FF}_{\mathrm{base}}$ applies to Standard and QK-Norm; $\mathrm{FF}_{\mathrm{MAP}}$ is the reduced QK-MAP feed-forward width used for parameter matching. All runs use sequence length 2048 and global batch size 2,097,152 tokens.

B Experimental Setup

We compare Standard attention, QK-Norm, and QK-MAP at 190M, 370M, 760M, 1B, and 7B parameters. The five scales are trained for 2B, 4B, 8B, 10B, and 20B tokens, respectively. All runs use sequence length 2048 and a global batch size of 2,097,152 tokens.

Within each scale, the three variants share the layer count, hidden size, attention-head configuration, tokenizer, pretraining mixture, optimizer, token budget, global batch size, initialization seed, and data order. The pretraining corpus is the same Dolma-based OLMo mixture for every variant.

QK-MAP adds a periodic projection. We reduce its feed-forward width so that the total parameter count remains closely matched to Standard attention. The difference is at most 0.033% at any scale and 0.017% at 7B.

The 190M architecture metadata were reconstructed from the archived training code and records.

B.1 Downstream Evaluation

Avg-8 is the unweighted mean accuracy over ARC-Easy, PIQA, SciQ, ARC-Challenge, COPA, HellaSwag, OpenBookQA, and WinoGrande. All eight tasks are evaluated zero-shot.

The common evaluation checkpoints for 190M, 370M, 760M, 1B, and 7B are steps 954, 1908, 3815, 4769, and 9465, respectively. All variants within a scale use the same checkpoint and evaluation settings.

B.2 Nearby-Checkpoint Evaluation

We also evaluate 16 common 7B checkpoints from steps 9455 to 9537, shown in Figure 4(c). QK-MAP ranks first at every checkpoint. Its Avg-8 margin over Standard averages 0.45 percentage points and ranges from 0.06 to 0.90 points. The consistent ordering across nearby endpoints shows that the result is not caused by selecting one favorable checkpoint.

C Detailed Configurations and Reproducibility

This section gives the scale-dependent architectures, parameter matching, shared optimization settings, and exact evaluation protocol needed to reproduce the experiments.

C.1 Scale-Specific Configurations and Parameter Matching

Values for 370M–7B come from exact resolved run configurations. The 190M metadata were reconstructed from archived training code.

Tables 4 and 5 isolate the attention formulation while keeping the backbone, training budget, and total parameter count closely aligned. The largest parameter difference from Standard is 0.033%, and the difference at the largest scale is 0.017%. Thus, a material increase in model size is unlikely to explain the reported optimization and downstream differences.

Scale	Standard	QK-Norm	QK-MAP	Δ params vs. Standard
190M	267,406,080	267,424,512	267,419,904	+0.0052%
370M	473,990,144	474,022,912	474,145,792	+0.0328%
760M	912,311,808	912,360,960	912,348,672	+0.0040%
1B	1,619,077,120	1,619,150,848	1,619,279,872	+0.0125%
7B	7,298,355,200	7,298,617,344	7,299,600,384	+0.0171%

Table 5: Analytical parameter counts from the resolved 370M–7B configurations and reconstructed 190M metadata. The Q/K RMSNorm and periodic-map parameters produce small differences. QK-MAP remains within 0.033% of Standard at every reported scale; the final column reports its relative parameter increase.

Field	Value	Field	Value
Optimizer	AdamW	RoPE theta	500,000
Betas / epsilon	$(0.9, 0.95)$ / 10^{-8}	Tokenizer	`allenai/dolma2-tokenizer`
Weight decay / grad clip	0.1 / 1.0	Vocabulary / padded size	100,278 / 100,352
Warmup / decay	5% / cosine	Data mixture	`OLMo-mix-0625-20Bsample`
Final LR ratio	0.1	Experiment / model / data seeds	12536 / 0 / 34521
Z-loss multiplier	10^{-5}	QK-Norm epsilon	10^{-6}
Parameter / reduction dtype	BF16 / FP32	QK-MAP periodic ratio	0.5
QK-MAP residual init	per-dimension, zero	QK-MAP map scope	shared by Q and K

Table 6: Settings shared across the reported scales and variants.

Task	Metric	N	Task	Metric	N
ARC-Easy	Acc.-unc.	570	ARC-Challenge	Acc.-unc.	299
PIQA	Acc./token	1000	COPA	Accuracy	100
SciQ	Accuracy	1000	OpenBookQA	Acc.-unc.	500
HellaSwag	Acc./token	1000	WinoGrande	Acc./token	1000

Table 7: Downstream evaluation protocol. All tasks use the validation split and zero-shot prompting. Acc.-unc. denotes unconditioned accuracy; Acc./token denotes length-normalized accuracy.

C.2 Shared Optimization and Implementation

Together, the shared settings in Table 6 make each within-scale comparison matched in optimizer, precision, tokenizer, data order, and initialization seed. This pairing isolates the attention formulation while keeping the main sources of training variation fixed.

C.3 Released Implementation and Compute

The code package contains the QK-MAP module, OLMo integration helpers, training configuration code, magnitude-response and norm-shuffling probes, and OLMES aggregation and bootstrap scripts. The implementation documents tensor shapes, initialization, residual options, and the order of Q/K projection and RoPE. Pretraining was run with eight NVIDIA A100 40GB GPUs using distributed data parallelism. The same software stack and precision settings were used for all variants within each scale.

D Evaluation Protocol

All tasks are evaluated zero-shot. The random-subsampling seed is 1234. The maximum evaluation size is 1000 examples, except that all 500 OpenBookQA validation examples are used.

Table 7 fixes the split, prompting, subsampling, and metric for the eight downstream tasks before comparing model variants. Avg-8 averages the eight task accuracies without weighting by dataset size. This gives equal weight to each capability measured by the benchmark suite.

Scale	Variant	ARC-E	PIQA	SciQ	ARC-C	COPA	HS	OBQA	WG	Avg-8
370M	Standard	39.12	58.90	67.00	**31.10**	60.00	25.90	28.80	52.00	45.35
	QK-Norm	40.70	59.40	70.00	23.75	**64.00**	27.00	33.60	48.80	45.91
	QK-Clip	38.25	59.00	68.00	29.10	**64.00**	27.30	30.40	47.60	45.46
	QK-MAP-D	40.70	**60.30**	71.20	28.09	61.00	**27.50**	31.40	**52.70**	**46.61**
	QK-MAP-P	**41.05**	58.80	**72.10**	25.75	61.00	27.20	33.80	48.90	46.08
	QK-MAP-R	40.70	58.00	69.70	25.42	61.00	27.30	**35.60**	49.80	45.94
760M	Standard	47.72	62.00	79.80	**31.44**	64.00	32.30	34.40	50.30	50.24
	QK-Norm	47.19	61.70	77.50	29.77	**65.00**	32.30	33.60	49.80	49.61
	QK-MAP-D	45.09	**63.30**	**81.00**	30.10	64.00	32.90	**38.60**	51.30	**50.79**
	QK-MAP-P	**47.89**	63.00	80.20	31.10	63.00	32.20	36.40	**51.60**	50.67
	QK-MAP-R	45.09	62.50	79.90	28.43	64.00	**33.20**	34.80	49.90	49.73
1B	Standard	48.42	**63.90**	**82.00**	27.76	66.00	35.00	36.00	**51.70**	51.35
	QK-Norm	**51.05**	63.40	79.40	31.77	66.00	35.70	**37.40**	51.50	52.03
	QK-MAP-D	47.72	**63.90**	80.10	32.78	66.00	33.30	34.60	50.10	51.06
	QK-MAP-P	50.18	63.50	80.60	31.44	**67.00**	**36.40**	35.60	50.00	51.84
	QK-MAP-R	49.30	**63.90**	**82.00**	**34.78**	**67.00**	34.20	34.40	51.30	**52.11**

Table 8: Downstream architecture ablations at the matched checkpoint for each scale. All entries are accuracies (%), and Avg-8 is computed from the eight displayed tasks. QK-MAP-D and QK-MAP-P omit the residual, whereas QK-MAP-R is the parameter-matched residual implementation used as the default QK-MAP design. QK-Clip was evaluated only at 370M. Best values within each scale are bold.

For validation CE, Dolma-6 is the mean over Books, Common Crawl, PeS2o, Reddit, Stack, and Wiki. The main table's Val. column averages C4, Pile, M2D2-S2ORC, and Dolma-6 with equal weight. All methods use the same validation shards and evaluator within each scale.

E Architecture and Residual Ablations

QK-MAP-D uses a same-dimensional periodic map without a residual. QK-MAP-P adjusts the map to match parameters without the residual. QK-MAP-R is the parameter-matched map with a learned per-dimension residual used as the default QK-MAP design in the cross-scale experiments. At 370M, we also compare against QK-Clip, the optimizer-side logit-control technique introduced in Kimi K2 (Kimi Team et al., 2025). All averages below are computed from the task-level evaluation outputs.

Table 8 compares these variants at 370M, 760M, and 1B. At each scale, at least one QK-MAP variant outperforms both Standard attention and QK-Norm. The direct map QK-MAP-D leads at 370M and 760M, while the parameter-matched residual map QK-MAP-R leads among periodic variants at 1B. This heterogeneous ablation shows that the gain is not tied to a single integration rule. It also supports the residual design at larger scale, where the periodic branch can refine the original representation.

F Attention Diagnostics

F.1 Synthetic Evidence Contrast

The main comparison in Table 2 averages the last-query evidence attention SNR over all layers. The query is the final question token; inserted evidence and the surrounding distractor tokens define the signal and noise regions. The same masks and layer-pooling rule are used for all variants.

The single-needle set contains 112,200 examples: 600 examples per setting, eleven depth layouts, and seventeen sequence lengths from 32 to 600 tokens. The two- and three-needle sets contain 201,600 and 162,000 examples, respectively, and the mixed-task set contains 112,200 examples. All sets use seed 20251231 and the 7B step-9465 checkpoints.

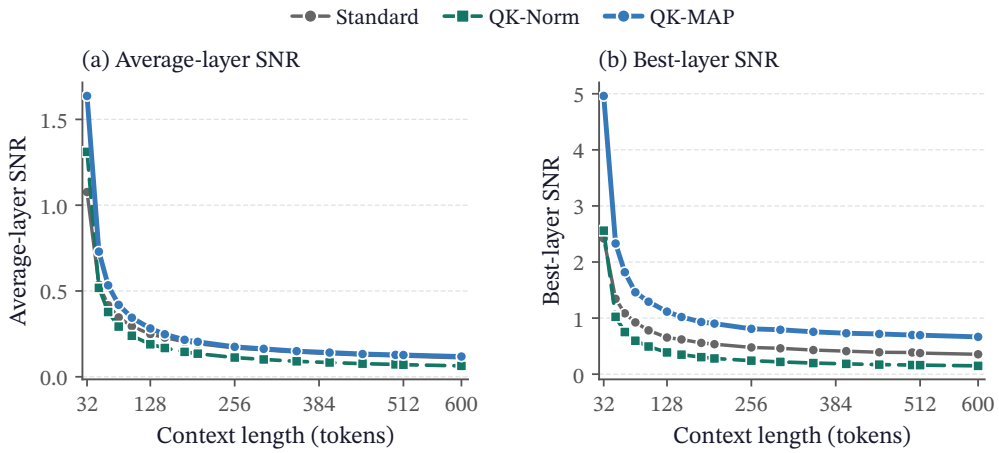

Figure 7: Single-needle attention SNR under two layer-pooling rules. (a) Average across layers. (b) Best layer for each example. Each point aggregates 6,600 examples.

Figure 7 compares two layer-pooling rules. Average-layer SNR uses every layer for each example. Best-layer SNR takes the largest layer score within each example before averaging over examples. The main text uses the average-layer result, and the best-layer view shows that the ordering also holds for the strongest layer response. QK-MAP leads both baselines at every evaluated length under both pooling rules.

F.2 Cross-Domain First-Token Allocation

For the 7B checkpoint, we measure the percentage of attention assigned to the first key, averaged over layers, heads, and query positions:

$$\text{F-Attn} = \frac{100}{LHT} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \sum_{i=1}^{T} A_h^{\ell}[i, 1]. \tag{40}$$

The statistic is computed on consecutive full-length validation chunks; incomplete final chunks are discarded. Uniform causal attention has reference value $100\mathcal{H}_T/T$, where $\mathcal{H}_T$ is the Tth harmonic number.

T	Variant	C4	Books	CC	PeS2o	Reddit	Stack	Wiki	ICE	M2D2	Pile	W103	Avg. ↓
32	Standard	21.91	21.36	21.71	23.79	21.89	22.39	22.71	20.86	22.65	22.97	25.54	22.53
	QK-Norm	19.84	20.05	19.68	21.19	20.58	22.59	20.25	20.58	21.07	21.51	24.83	21.11
	QK-MAP	19.80	19.06	19.57	21.02	20.35	20.06	19.97	19.61	20.95	20.56	24.43	**20.49**
64	Standard	14.09	14.49	13.88	15.73	13.87	15.27	14.67	13.77	14.34	15.17	17.93	14.84
	QK-Norm	13.20	14.07	12.88	14.23	13.32	16.25	13.37	14.17	13.82	14.80	17.26	14.31
	QK-MAP	12.35	12.75	12.16	13.56	12.79	13.73	12.68	12.91	12.96	13.25	17.01	**13.29**
128	Standard	8.90	9.96	8.77	10.23	8.46	10.52	9.26	9.02	8.47	9.93	11.96	9.59
	QK-Norm	8.83	9.82	8.64	9.80	8.62	11.90	8.93	9.83	8.77	10.25	11.63	9.73
	QK-MAP	7.64	8.63	7.54	8.65	7.73	9.57	7.82	8.50	7.53	8.54	11.41	**8.50**
256	Standard	5.52	7.03	5.47	6.59	4.92	7.26	5.80	5.86	4.73	6.51	7.62	6.12
	QK-Norm	5.81	6.98	5.66	6.55	5.47	8.58	5.95	6.72	5.33	7.03	7.64	6.52
	QK-MAP	4.66	6.12	4.52	5.41	4.57	6.68	4.79	5.55	4.21	5.52	7.42	**5.40**
512	Standard	3.40	4.94	3.26	4.13	2.74	4.97	3.66	3.77	2.55	4.21	4.68	3.85
	QK-Norm	3.84	4.76	3.46	4.25	3.24	5.85	3.91	4.48	3.04	4.71	4.79	4.21
	QK-MAP	2.89	4.16	2.58	3.23	2.56	4.53	2.98	3.59	2.29	3.57	4.62	**3.36**

Table 9: F-Attn (%) over eleven validation domains for the 7B step-9465 checkpoint. Lower values indicate less average attention assigned to the first key; averages are unweighted over domains. Books = Dolma Books; CC = Dolma Common Crawl; PeS2o = Dolma PeS2o; Stack = Dolma Stack; Wiki = Dolma Wiki; M2D2 = M2D2-S2ORC; and W103 = WikiText-103.

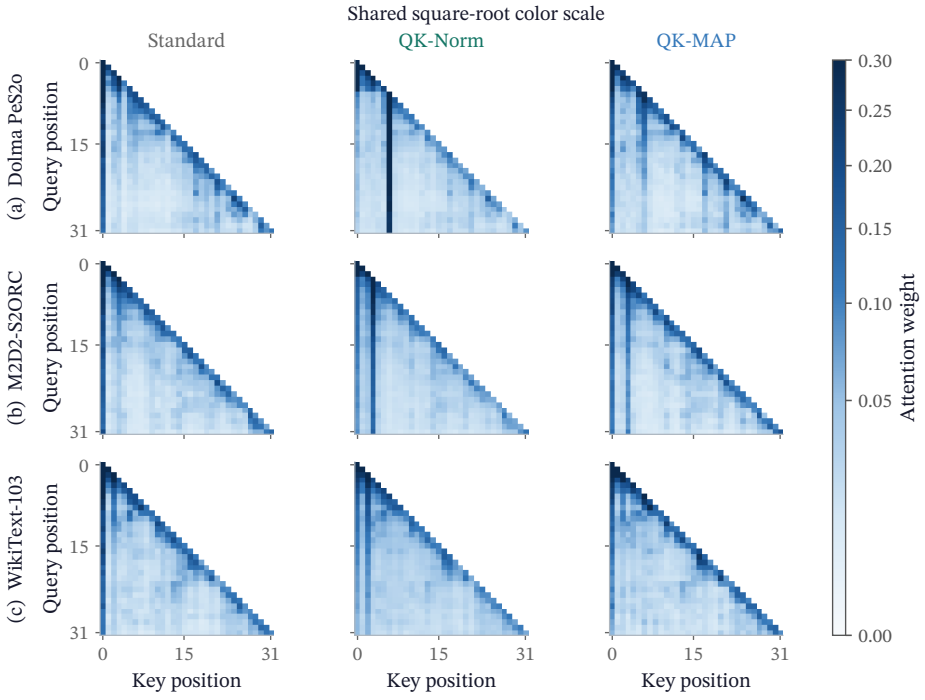

Figure 8: Complete causal-attention matrix display for WikiText-103, M2D2-S2ORC, and Dolma PeS2o at the 7B step-9465 checkpoint. The nine layer/head-averaged matrices are shown without domain-level selection and share one square-root color scale. Across all three domains, QK-MAP shows a weaker first-column band, matching the lower first-token allocation reported in Table 9. Across all 55 domain–length cells, QK-MAP has lower F-Attn than both baselines; its eleven-domain mean is 9.1–12.7% lower than Standard across lengths 32–512.

Norm scope	Variant	Shuffle Q	Shuffle K	Shuffle Q and K
All heads	Standard	0.01651	0.18816	0.22541
	QK-Norm	0.00001	0.00012	0.00013
	QK-MAP	0.00617	0.19360	0.20106
Per head	Standard	0.17644	0.76922	1.14959
	QK-Norm	0.10227	0.62821	0.77627
	QK-MAP	0.01829	0.32990	0.37760

Table 10: Mean ΔCE for global and headwise norm shuffling. The global scope matches the QK-RMSNorm axis.

Variant	Shuffle Q	Shuffle K	Shuffle Q and K
Standard	[0.01503, 0.01802]	[0.18135, 0.19568]	[0.21772, 0.23350]
QK-Norm	[−0.00003, 0.00005]	[−0.00009, 0.00033]	[−0.00010, 0.00034]
QK-MAP	[0.00512, 0.00736]	[0.17708, 0.21414]	[0.18283, 0.22208]

Table 11: Paired 95% bootstrap intervals for global norm shuffling.

G Magnitude Intervention and Text Perturbation

G.1 Norm-Shuffling Protocol

We use three trained 190M checkpoints at approximately 2B tokens: Standard, QK-RMSNorm, and the same-dimensional direct map QK-MAP-D. These probes use the direct map rather than the parameter-matched residual variant in Table 1. The evaluation sample is the first 128 consecutive 512-token chunks of the C4 validation token shard, evaluated in FP32 with batch size 16.

Intervention. The hook acts after Q/K normalization and RoPE, immediately before the attention backend, in all twelve layers. Let $v_{b,t}$ denote the query or key vector for sequence b at position t. For a permutation π_t over the batch, we replace it by

$$\widetilde{v}_{b,t} = \|v_{\pi_t(b),t}\|_2 \frac{v_{b,t}}{\|v_{b,t}\|_2}. \tag{41}$$

The implementation clamps denominators below 10^{-12}. The permutation is restricted to the same position across sequences, so the intervention never moves a future-position norm to an earlier position. Query-only and key-only conditions change one side; the joint condition shuffles both. The main result uses the norm of the full 768-dimensional vector formed by concatenating twelve heads. This matches the QK-RMSNorm axis in these checkpoints.

Measurement and controls. We report the change in each sequence's CE relative to its unmodified forward pass, then average across sequences and three permutations. The seed is 20260725. We obtain paired 95% bootstrap intervals by resampling the 128 sequences 10,000 times after averaging repetitions. The intervals describe variation across this evaluation sample. The no-hook, identity-hook, and split-and-reconstruct controls agree to within 9.54×10^{-7} CE per sequence. This probe evaluates raw contiguous chunks; its clean CE serves only as the reference for the paired change, rather than replacing the standard validation evaluator.

Normalization axis. Table 10 also reports shuffling each head's norm separately. Global normalization removes overall token scale but retains variation in how magnitude is allocated among heads. Accordingly, QK-RMSNorm is almost unchanged by global shuffling but responds to headwise shuffling. The main comparison targets the global scale that normalization directly suppresses. The intervention measures the use of projected query/key magnitudes; it does not isolate which part of the periodic representation produced that variation.

G.2 WORD-REPLACEMENT PROTOCOL

We evaluate Standard, QK-RMSNorm, QUEST, and QK-MAP-D at 190M after approximately 2B pretraining tokens. We decode the archived WikiText-103 validation token stream, retain the text through the first 20,000 alphabetic words, and re-tokenize each perturbed version with the original training tokenizer, preserving its saved pre-tokenization rules. For word index j, we generate a fixed uniform draw using seed $20260926 + j$ and replace that word by `AAA` if the draw is below the chosen rate. Special tokens are left unchanged. All models receive identical text. The replacement sets are nested as the rate increases.

The rates 1.5%, 2.5%, and 5% replace 312, 525, and 1,023 words, respectively. We score every next-token target in causal blocks of at most 512 targets, including the shorter final block. The clean and perturbed versions contain 28,280, 28,247, 28,220, and 28,156 scored targets. Perplexity is $\exp(\sum_t \mathrm{CE}_t/N)$ over all scored targets. This fixed-block implementation differs from QUEST's sliding-window protocol, so Table 3 is a matched comparison among our checkpoints, not a reproduction of the original paper's absolute perplexities.

QK-MAP has the lowest absolute perplexity at every rate. At 5% replacement, the increases relative to each model's own clean perplexity are 56.1%, 59.0%, 58.7%, and 60.3% for Standard, QK-RMSNorm, QUEST, and QK-MAP, respectively. The main conclusion therefore concerns the retained absolute language-modeling advantage under word replacement.