SCIENCE OF LONG CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models increasingly operate over long inputs, such as long documents, code repositories, and their own reasoning traces, and must retrieve information from any position in them. However, language model performance degrades as input length increases, and the extent of the degradation depends on the architecture. Every architectural choice is an inductive bias with its own limitations, and these limitations are poorly understood. We study them by training many architectures from scratch in a controlled setting, where the models differ only in architecture and serve as controls for each other. We find that removing explicit positional encoding alone does not improve length generalization, and that RoPE's retrieval beyond the training length degrades due to keys at offsets never observed during training. Hybrid models that restrict most layers to a sliding window still read far keys at these offsets in their global attention layers. We then show that in a relative positional encoding, keeping positions within the trained range comes at the cost of resolving nearby positions. Building on this insight, we design Log-Ratio RoPE, which rotates queries and keys by their log-position to keep offsets close to the trained range. We use it in the global attention layers of a hybrid model, while sliding-window layers retain RoPE to resolve nearby positions within the trained range. In 0.5B models trained at 8K, this design improves accuracy on the RULER long-context benchmark at 64K, eight times the training length, by 18.7 percentage points over RoPE and 8.1 percentage points over the RoPE–NoPE hybrid.

1 INTRODUCTION

Alongside scaling training compute (Kaplan et al., 2020; Hoffmann et al., 2022), language models are now also improved by scaling test-time compute: reasoning models produce long chains of thought before answering (Guo et al., 2025; OpenAI et al., 2026). Harder tasks often elicit longer reasoning traces, so models must operate reliably over increasingly long contexts. The ability to retrieve and use information from any position in a long context is therefore a requirement for scaling test-time compute. Long-context ability also benefits other applications, such as answering questions over long documents (Bai et al., 2024) and working on large code repositories (Jimenez et al., 2024). However, model performance degrades as input length increases (Hong et al., 2025; Hsieh et al., 2024; Liu et al., 2023). The degradation is sharpest beyond the training length, and generalizing beyond it (*length generalization*) matters because training on longer sequences is costly and long documents are scarce.

A model's ability to use long contexts depends on its architecture (Sec. 2). Throughout this work, we use *architecture* to refer to the choice of positional encoding and to how global and local attention are composed across the model. In *global attention*, each token attends to all previous tokens. In *local attention* (also called *sliding-window attention*; Beltagy et al., 2020), each token attends only to a fixed window of recent tokens. Prior work has shown that causal transformers can learn positional information even without explicit positional encoding (NoPE; Haviv et al., 2022; Kazemnejad et al., 2023). Explicit positional encoding is therefore not strictly necessary for sequence modeling. A relative positional encoding, such as RoPE (Su et al., 2023), makes the attention logit depend only on a relation between the query and key positions. In RoPE, this relation is their difference, the offset $\Delta = m - n$ for query position m and key position n. While explicit positional encoding can improve sequence modeling, its benefits are not uniform (Sec. 2). More generally, every architectural choice, including positional encoding and how global and local attention are composed, is an inductive bias that makes some patterns easier to learn but also introduces limitations.

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.

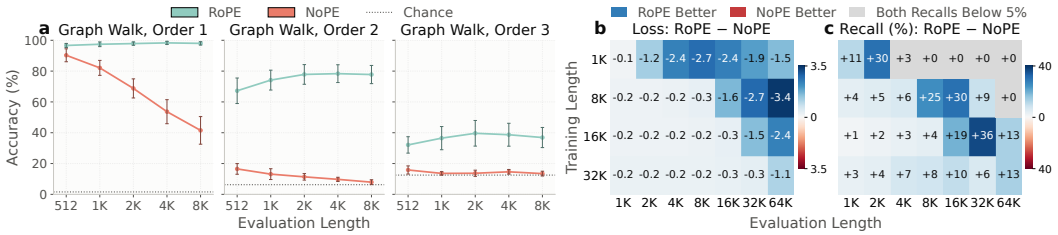

Figure 1: **(a)** Graph walk accuracy for RoPE and NoPE trained at 8K, one panel per Markov order c. The dotted line marks chance. Error bars are 95% confidence intervals over 8 graphs. **(b, c)** RoPE minus NoPE in log PPL (b) and key-value recall (c) for each training and evaluation length. Blue cells favor RoPE. Grey cells mark recall below 5% for both models.

These limitations are poorly understood. The goal of this paper is to better understand the limitations of these inductive biases and to leverage that understanding to improve long-context performance. We compare many architectures in a single controlled setting. RoPE, NoPE, ALiBi (Press et al., 2022), sliding-window attention, and local–global hybrids are all trained from scratch with the same data, total number of training tokens, and optimizer, so that only the architecture varies. This allows one architecture to serve as a control for another. For example, NoPE has no offsets, so it shows how a model fails when only the context length grows (Sec. 3), and two hybrids that differ only in their global layers isolate the effect of the positional encoding in those layers (Sec. 4). We also vary the training length and the head dimension, and train each 64M-parameter model at three learning rates. We then validate our final design at 0.5B parameters (Sec. 8).

Contributions. We show that removing explicit positional encoding alone does not improve length generalization (Sec. 2), and that RoPE's retrieval beyond the training length degrades because of keys at offsets never observed in training (Sec. 3). Restricting most layers to a sliding window does not remove this failure: global attention still reads far keys at these offsets, where RoPE is worse for retrieval than no positional encoding (Sec. 4). This explains why the RoPE–NoPE hybrid, one of the best-performing long-context architectures (Yang et al., 2025a) and used in recent models (Cohere et al., 2025; Meta Superintelligence Lab, 2026), outperforms the RoPE–RoPE hybrid. We prove that no relative positional encoding can keep almost all keys in the trained range (the range of Δ seen in training) while also keeping its positional resolution, the ability of a query to tell two keys apart by their positions (Theorem 1). A hybrid can split the two properties across layers: RoPE in the local layers provides positional resolution, and Log-Ratio RoPE, which replaces $\Delta = m - n$ with $\Delta_{\ln} = \ln m - \ln n$, keeps almost all keys of the global layers in the trained range (Sec. 6). At 0.5B parameters, this hybrid matches RoPE on short-context benchmarks and improves RULER accuracy at 64K, eight times the training length, by 18.7 points over RoPE and 8.1 points over the RoPE–NoPE hybrid (Sec. 8). These results show that RoPE helps at trained offsets but hurts at unseen ones, so it should be restricted to layers whose offsets stay within the trained range.

2 What Explicit Positional Information Adds

A causal transformer with no positional encoding (NoPE) receives no explicit information about token positions. Nevertheless, the causal mask induces position-dependent attention contexts: a query at position m can attend to exactly m keys. Across layers, the model can use this structural asymmetry to infer position (Chi et al., 2023). We therefore use NoPE as the baseline: comparing a positional encoding against NoPE isolates what its added positional information contributes.

Positional Encoding Facilitates Learning Abstractions. Inferring position is an additional learning requirement for a NoPE model, and it may limit what else the model learns. To isolate this effect from language modeling, we compare the models on a fully synthetic graph-walk ICL task (Sec. B.3). The tokens are arbitrary symbols for graph states, reassigned for every graph, so no transition can be memorized: the model must predict the next token with a rule that holds for any symbols. We call such a rule an abstraction. A model that learns abstractions can handle patterns

it never saw in training from the context alone, which is what in-context learning requires (Brown et al., 2020). Applying this rule requires knowing which token came after which. We vary the Markov order c, the number of last tokens that fix the next token (Edelman et al., 2024). Order 1 is the rule implemented by an induction head (Olsson et al., 2022), and orders 2 and 3 require matching the last two or three tokens (Akyürek et al., 2024). The evidence is always local, so the task tests rule learning rather than long-range retrieval. Neither model is trained on the task.

At order 1, RoPE stays above 95% accuracy throughout the training window, while NoPE declines from 90% to approximately 40% (Fig. 1a). Inside the training window, RoPE reaches approximately 75% at order 2 and 40% at order 3. NoPE stays near chance on both. Thus, at the scale we study, RoPE learns ICL rules that NoPE does not. This result does not imply that NoPE cannot learn these rules at a larger scale.

2.1 Context Utilization and Length Generalization

During pre-training, a model is exposed to every position in its training context, but next-token prediction (NTP) does not require the model to use all of that context. NTP loss can often be minimized with recent context, so a model trained with a large S can reach low loss without learning to retrieve information from distant tokens (Fig. 6) (Fang et al., 2025). We separate two challenges: using long context within the training length (*context utilization*), and length generalization beyond it. To measure retrieval, we report key-value recall (Sec. B.2) alongside loss. The task places 8 key-value pairs in a document and requires the model to find an earlier occurrence of a queried key and copy the token that followed it.

Inductive Biases Can Limit Context Utilization. Explicit positional information is not sufficient for context utilization. ALiBi adds a distance-dependent bias to the attention logits, and the sliding-window attention (SWA) model uses local attention in every layer. Both constraints limit what the model learns from distant tokens, and both models have lower key-value recall than NoPE even within their training context. Sec. D.1 gives the full analysis.

NoPE Does Not Improve Extrapolation. A common argument against explicit positional encoding is that it impairs length generalization: Kazemnejad et al. (2023) found that NoPE extrapolates better than positional encodings on small synthetic reasoning tasks. We test whether this advantage persists in language modeling, and whether NoPE's extrapolation improves with longer training contexts, by training NoPE and RoPE models at context lengths $S \in \{1\text{K}, 8\text{K}, 16\text{K}, 32\text{K}\}$ (Appendix A) and comparing models trained at the same S (Fig. 1b,c).

Increasing S extends NoPE's usable context. Inside the training window, NoPE already has higher loss than RoPE, with the largest gap near S, but its recall remains comparable to RoPE's up to approximately $S/2$. Beyond the training length, the gap widens sharply. At $2S$, NoPE's loss exceeds RoPE's by more than one nat for every training length, and for training lengths up to 16K NoPE recall falls to chance while RoPE still answers approximately one third of queries correctly. Together with the graph-walk results, removing positional encoding discards useful structure without improving extrapolation. RoPE uses its training context best among these architectures, yet it also degrades beyond the training length. We next study why (Sec. 3).

3 Why RoPE Fails Beyond the Training Length

Rotary position embeddings (RoPE) (Su et al., 2023) split each query and key into two-dimensional pairs. Pair j is rotated at frequency $\theta_j = b^{-2j/d}$, where b is the RoPE base and d is the head dimension. The attention logit is

$$\big\langle \mathrm{R}(m)\, q,\ \mathrm{R}(n)\, k \big\rangle = q^\top \mathrm{R}(-\Delta)\, k = \sum_j |q_j|\,|k_j| \cos(\phi_j + \Delta\theta_j), \tag{1}$$

where $\Delta = m - n$ is the offset between query position m and key position n. The matrix $\mathrm{R}(m)$ rotates pair j by the angle $m\theta_j$, and ϕ_j is the signed angle from the pair k_j to the pair q_j. A model trained on contexts of length S observes only offsets $0 \le \Delta < S$. Prior work attributes the failure past the training length to RoPE's slow frequencies. A frequency whose period exceeds S completes

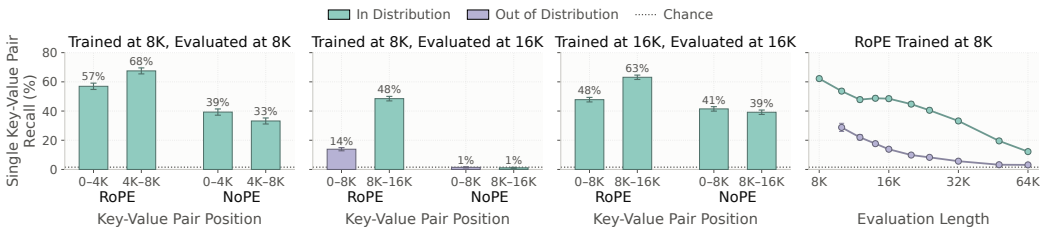

Figure 2: Recall of a single key-value pair moved through the context (Sec. C.1). **First three:** recall averaged over each half of the context. A half is in distribution when all its offsets to the query are below the training length. **Fourth:** the RoPE model trained at 8K, with recall averaged over trained and OOD offsets at each evaluation length. Error bars are 95% confidence intervals.

less than one rotation over the trained offsets, so longer contexts expose the model to phases that were not observed during training (Men et al., 2024; Liu et al., 2024). Context-extension methods rescale the frequencies or the positions so that a chosen target length maps back into the trained range (Chen et al., 2023; Peng et al., 2026), but a context longer than the target again reaches unseen offsets (Lu et al., 2024). A larger head dimension does not remove the failure either: RoPE models with head dimension up to 256 and training length up to 32K all degrade sharply past their training length. Sec. D.2 gives the full analysis. We instead examine the cost of the unseen offsets and address it in the positional encoding itself (Sec. 6).

OOD offsets limit extrapolation. Past S the model sees two changes: the softmax normalizes over $L > S$ keys, and keys at offsets $S \leq \Delta < L$ enter Eq. 1 at rotations never seen in training. Training fits the logit only on $\Delta < S$, so its dependence on position is not learned at larger offsets (Sec. 5.2). For a query at the end of the context, the S most recent keys stay at trained offsets and the $L - S$ earliest keys are at out-of-distribution (OOD) offsets. Keys at trained and OOD offsets compete in the same softmax over L keys. Comparing the retrieval of a key at a trained offset with that of a key at an OOD offset therefore holds the number of keys fixed and changes only Δ. Retrieval also falls with Δ within the training length, so we compare this gap with the gap in a model trained at length L, which has the same offsets and the same number of keys, all seen in training.

We place a single key-value pair in a document, move it through the context, and query its key at the end. We compare the RoPE model trained at 8K and evaluated at 16K with the model trained at 16K (Fig. 2). At 16K, the model trained at 8K recalls 48% of pairs at trained offsets and 14% at OOD offsets, a gap of 34 points. The model trained at 16K recalls 63% and 48% at the same offsets, a gap of 15 points. The number of keys is the same for the four RoPE bars in the second and third panels of Fig. 2, so the additional 19 points lost by the model trained at 8K come from the offsets it did not observe in training. The OOD keys also lower recall at trained offsets: the model trained at 8K drops from 62% at 8K to 48% at 16K, while the model trained at 16K recalls 63% at the same offsets with the same number of keys. Recall at trained offsets falls further as more keys lie at OOD offsets, to 33% at 32K. The NoPE model has no offsets: trained at 16K, it recalls 41% and 39% in the two halves, so the fall of recall with Δ within the training length is specific to RoPE. The same result holds with 8 key-value pairs at fixed positions (Sec. C.1).

4 Hybrid Models

Hybrid architectures interleave SWA layers with global attention layers. They were originally motivated by efficiency: local attention reduces the cost of processing long sequences while periodic global attention preserves access to the full context (Beltagy et al., 2020; Team et al., 2024; 2025). The failure mode in Sec. 3 suggests a simple way to improve length extrapolation. If most layers attend only within a fixed local window, their offsets remain within the trained range regardless of the total sequence length. Several recent hybrids use a 3:1 ratio of local to global attention (Yang et al., 2025a; Cohere et al., 2025), so only one quarter of the attention layers are exposed to OOD offsets. We test whether this is sufficient using three 3:1 hybrids trained at 8K, denoted by the positional encoding of their local and global attention: RoPE–RoPE, NoPE–NoPE, and RoPE–NoPE (Fig. 3).

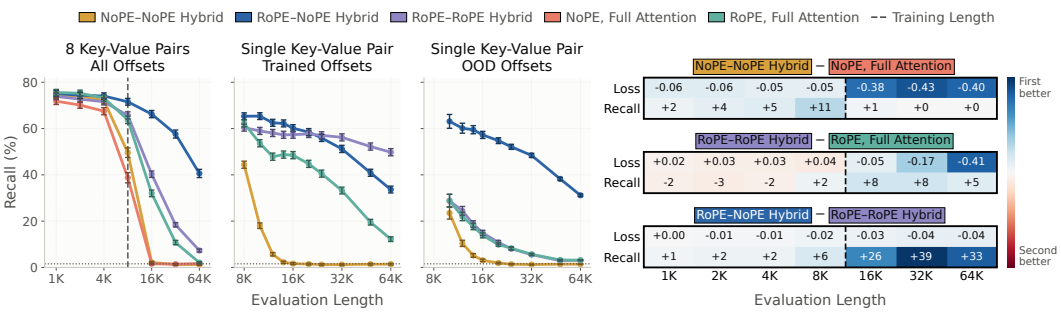

Figure 3: Three 3:1 hybrids and both full-attention models, trained at 8K. **Left:** Recall of 8 key-value pairs. **Middle two:** Recall of a single key-value pair at trained (<8K) and OOD (>8K) offsets, as in Fig. 2. **Right:** Each hybrid minus its full-attention model (top two) and RoPE–NoPE minus RoPE–RoPE (bottom), in loss and in recall of 8 key-value pairs. Blue favors the first model. The dashed line marks the training length.

Limiting out-of-distribution exposure is insufficient. If limiting OOD exposure were sufficient, each hybrid would extrapolate better than its full-attention model. The hybrid layout does not eliminate the NoPE failure at the training length (Fig. 3, left). Like NoPE full attention, the NoPE–NoPE hybrid's loss rises by 1.1 nats from 8K to 16K and its recall of 8 key-value pairs falls to chance by 16K, even though only its global attention attends beyond the local window. As in Sec. 2, the RoPE–RoPE hybrid does not exhibit this failure and has lower loss and higher recall beyond the training length. The top two heatmaps compare each hybrid with its full-attention model. Beyond the training length, the hybrid layout improves RoPE recall by 5 to 8 points and changes NoPE recall by at most 1 point. The small gain for RoPE is consistent with Sec. 3. A key more than S tokens from the query lies outside every local window, so only global attention can read it, and it reads the key at an OOD offset. The middle panels confirm this. At OOD offsets, the RoPE–RoPE hybrid's recall of a single key-value pair is within 3 points of RoPE full attention at every length. Its gain is at trained offsets, where it keeps 50% recall at 64K, against 12% for RoPE full attention. The hybrid layout therefore reduces the number of layers that see OOD offsets, but not the OOD offsets on which long-range retrieval depends.

RoPE in the global attention degrades retrieval. To test whether RoPE in the global attention affects extrapolation, we compare the RoPE–RoPE hybrid with the RoPE–NoPE hybrid. Both use RoPE in the same local attention, where offsets remain within the trained range, and therefore differ only in the global attention, which encounters OOD offsets: RoPE–NoPE has no positional encoding there, whereas RoPE–RoPE uses RoPE. If limiting RoPE's exposure to the global attention is sufficient, the two hybrids should extrapolate similarly. If RoPE itself fails once the global attention operates beyond the training length, the RoPE–RoPE hybrid should perform worse.

The bottom heatmap of Fig. 3 supports the second hypothesis. Within the training length, the two hybrids are within 6 points of each other in recall of 8 key-value pairs. Beyond that, the RoPE–NoPE hybrid achieves 26 to 39 points higher recall, while their losses differ by at most 0.04 nats. At 64K, the RoPE–NoPE hybrid reaches 41% recall, compared with 7% for the RoPE–RoPE hybrid. The middle panels show that this difference is at OOD offsets. There the RoPE–NoPE hybrid keeps 57% recall of a single key-value pair at 16K and 31% at 64K, close to its recall at trained offsets, while the RoPE–RoPE hybrid falls to 15% and 3%. This does not mean that NoPE extrapolates better: the NoPE–NoPE hybrid, which has no positional information in any layer, is the weakest hybrid. Rather, RoPE at OOD offsets is worse for retrieval than no positional encoding. Thus RoPE helps at trained offsets (Sec. 2) but hurts at unseen ones. This motivates a positional encoding for the global attention whose offsets remain close to the trained range (Sec. 5).

5 Ideal Positional Encoding

Prior work has proposed properties of an ideal positional encoding, such as a bias towards recent tokens or a logit that decays with distance, but these properties often add inductive biases that

are stronger than necessary (Sec. G). Every positional encoding carries some inductive bias, but an overly strong inductive bias can prevent the model from properly learning to use its context (Sec. D.1). At a minimum, a positional encoding should provide positional information and allow the model to learn how to use it during training. Beyond this, an ideal positional encoding should satisfy three properties. First, relativity: the encoding must convey only relative positional information. Second, length generalization: at any context length, the encoding must keep almost all keys within the range of relative positions seen in training. Third, positional resolution: the encoding must distinguish every pair of positions, and this ability should fall as slowly as possible as the context grows. Su et al. (2023) formalized relativity for the difference $m - n$. We state it for a general relation between positions and formalize the remaining two properties below. We then prove that length generalization and positional resolution are inversely related: the more keys an encoding keeps in the trained range, the lower the positional resolution it can have (Theorem 1).

5.1 Relativity

A relative encoding provides the model the relation Δ between two positions directly, so the model does not need to infer it. The relation Δ should enter the attention logit through the query–key inner product, so that the effect of relative position depends on the content of the query and key, and the model can learn which offsets are relevant for which tokens (Sec. 7). Let q_m be the encoded query at position m and k_n the encoded key at position n.

Definition 1 (Relativity)**.** *Let $\Delta(m, n) \in \mathbb{R}$ be a relation between positions with $\Delta(m, m) = 0$. An encoding satisfies* relativity *if the logit depends on the positions only through Δ,*

$$\langle q_m,\, k_n \rangle \;=\; q^\top \mathrm{R}\big(\Delta(m,n)\big)^\top k, \tag{2}$$

for a continuous positional kernel $\mathrm{R} : \mathbb{R} \to \mathbb{R}^{d\times d}$ that (i) preserves content, $\|\mathrm{R}(\Delta)\, q\| = \|q\|$ for all Δ and q, and (ii) is equivariant in the relation, $\mathrm{R}(\Delta + \delta) = \mathrm{R}(\delta)\,\mathrm{R}(\Delta)$ for all Δ, δ.

Condition (i) lets position rotate the content but not scale it, so the encoding cannot act as a content-free gain or decay. Condition (ii) makes the kernel act on content in the same way at every value of Δ. When $\Delta(m, n) = m-n$, we recover standard RoPE. Following Schenck et al. (2025), any R that satisfies Def. 1 has the form $\mathrm{R}(\Delta) = \exp(\Delta \mathrm{A})$ for a fixed skew-symmetric generator $\mathrm{A} \in \mathbb{R}^{d\times d}$ (Sec. E.1). Relativity therefore fixes R but leaves Δ free, so we analyze Δ and its relation to the remaining properties.

Compute Efficiency. RoPE encodes each query and key once from its own position, so all logits are a single matrix product during both training and inference. A relative encoding keeps this compute efficiency if and only if $\Delta(m, n) = f(m) - f(n)$ for some position map f (proof in Sec. E.2). For the rest of our analysis, we consider any relation Δ that satisfies relativity (Def. 1) and note when a relation is not compute efficient.

5.2 Length Generalization

A positional encoding alone cannot guarantee that a model generalizes beyond its training length, since the model can also fail for reasons unrelated to position. The encoding determines only the positional input to attention, which under relativity is Δ (Sec. 5.1). The relevant property of the encoding is therefore whether a longer context produces values of Δ outside the range observed during training. Consider a model trained at length S, and let $\Delta_S = \{\Delta(m, n) : 1 \le n \le m \le S\}$ be the set of values of Δ that occur during training. We call $[\min \Delta_S, \max \Delta_S]$ the trained range. At a length $L > S$, a query at position $m \in (S, L]$ attends to $m > S$ keys. The key at position n is in the interpolation regime if $\Delta(m, n)$ lies in the trained range, even if this value never occurs in training (Sec. E.4). Otherwise, the key is in the extrapolation regime, and the model receives $\mathrm{R}(\Delta)$ at a value of Δ outside the range observed during training.

The encoding therefore affects length generalization through the keys in the extrapolation regime. The number of these keys can grow with the context, since a query at position m attends to m keys, so we measure their share rather than their number. Let

$$\phi_S(m) \;=\; \frac{1}{m}\,\big|\{\, n \le m : \Delta(m,n) \notin [\min \Delta_S, \max \Delta_S]\,\}\big| \tag{3}$$

be the share of the keys of a query at position m that are in the extrapolation regime.

Definition 2 (Length generalization)**.** *A positional encoding* $\mathrm{R}(\Delta)$ *is* length generalizable *if* $\sup_m \phi_S(m) \to 0$ *as* $S \to \infty$.

In RoPE, $\Delta = m - n$ and the trained range is $[0, S-1]$. A query at position $m > S$ has its $m - S$ earliest keys at offsets $\Delta \geq S$, in the extrapolation regime, so $\phi_S(m) = 1 - S/m$. This share goes to 1 as m grows, so $\sup_m \phi_S(m) = 1$ for every S, and RoPE is not length generalizable. Now consider the ratio of the key and query positions, $\Delta = 1 - n/m$, whose trained range is $[0,\, 1 - 1/S]$. A key is in the extrapolation regime only if $n < m/S$, so $\phi_S(m) = (\lceil m/S \rceil - 1)/m < 1/S$ for every m. Hence $\sup_m \phi_S(m) \leq 1/S$, which goes to 0 as S grows, and the ratio is length generalizable. Its other keys may take unseen values of Δ, but these values lie in the trained range. However, the ratio is not compute efficient (Sec. E.2).

5.3 Positional Resolution

We define positional resolution to be the ability of a query to tell two keys apart by their positions. Two keys with different content are easy to tell apart, since a generic query is not orthogonal to their difference. In the worst case, the two keys k_i and k_j share one content vector k and differ only in position, which enters through R. We measure positional resolution by the largest logit gap that unit vectors q and k can produce between two such keys.

Definition 3 (Positional resolution)**.** *Fix a query at position* m*. The positional resolution between two of its keys at positions* i *and* j *is the largest logit gap between them over unit* q *and* k*,*

$$\rho_m(i,j) \;=\; \max_{\|q\|=\|k\|=1} \big(\langle q_m,\, k_i\rangle - \langle q_m,\, k_j\rangle\big) \;=\; \big\|\mathrm{I} - \mathrm{R}\big(\Delta(m,i) - \Delta(m,j)\big)\big\|, \tag{4}$$

where $\|\cdot\|$ *is the spectral norm (Sec. E.3). The positional resolution of the query is the smallest of these over all pairs of its keys,*

$$r_m \;=\; \min_{1 \leq i < j \leq m} \rho_m(i,j).$$

Let $\theta_{\max}$ be the fastest frequency of A, the largest $|\theta_j|$ among its eigenvalues $\pm i\theta_j$. Since $\mathrm{R}(\Delta) = \exp(\Delta \mathrm{A})$,

$$\rho_m(i,j) \;\leq\; \theta_{\max} \big|\Delta(m,i) - \Delta(m,j)\big|,$$

and this bound is tight up to a constant factor while the right-hand side is at most π (Sec. E.4). Positional resolution therefore degrades as two keys take closer values of Δ, no matter how Δ is defined. Sec. E.4 works through this for the difference and the ratio.

5.4 Inverse Relation

Positional resolution and length generalization are both set by Δ, and they pull in opposite directions. Length generalization asks that almost all keys of a query take values of Δ inside the trained range, whose width is fixed. Positional resolution asks that no two keys take close values of Δ. A fixed interval cannot hold a growing number of values that are all far apart.

Theorem 1 (Inverse relation)**.** *Let* $D(S) = \max \Delta_S - \min \Delta_S$ *be the width of the trained range. For every encoding that satisfies relativity, every training length* S*, and every query position* m *with* $(1 - \phi_S(m))\, m \geq 2$*,*

$$r_m \;\leq\; \frac{\theta_{\max}\, D(S)}{(1 - \phi_S(m))\, m - 1}. \tag{5}$$

Hence a length generalizable encoding has $r_m = O(1/m)$*: its positional resolution falls at least as fast as* $1/m$*. An encoding with* $m\, r_m$ *unbounded is not length generalizable (proof in Sec. E.5).*

6 Log-Ratio RoPE

Theorem 1 shows that no positional encoding retains both length generalization and positional resolution as the context grows. A hybrid model does not need both in every layer. In a local attention layer, a query attends to at most S keys, so RoPE never leaves its trained range and retains its high positional resolution. Only the global layers must length generalize, and the resolution they lose can be supplied by the local layers. This split is consistent with Sec. 4, where keeping RoPE only in the

local layers improves retrieval beyond the training length. We therefore change only the relation Δ used in the global layers:

$$\Delta_{\ln} = \ln \frac{m}{n} = \ln m - \ln n, \tag{6}$$

the log of the ratio of the two positions rather than their difference. We call the resulting encoding Log-Ratio RoPE. This relation has the two properties the global layers need. First, it is length generalizable (Def. 2). Its trained range is $[0, \ln S]$, and a key falls outside this range only when $n < m/S$. As for the ratio in Sec. 5.2, less than a $1/S$ share of the keys of any query is in the extrapolation regime. Using Log-Ratio RoPE in every layer would lose the efficiency of local attention, which does not limit retrieval (Sec. 4), and using it in the local layers would lose their resolution. Second, $\Delta_{\ln}$ is compute efficient, since it is a difference $f(m) - f(n)$ with $f = \ln(\cdot)$ (Prop. 1). $\Delta_{\ln}$ enters the logit through R, so the logit of Eq. 1 becomes $q^\top R(-\Delta_{\ln})\, k$ with the j-th pair rotated by the angle

$$\alpha_j = \theta_j \Delta_{\ln} = \theta_j \big(\ln m - \ln n \big). \tag{7}$$

We set the frequencies to a geometric progression with base $b = 32$ (ablated in Sec. F.3) and set the highest frequency θ_0 from a large design horizon H (Sec. F.1):

$$\theta_j = \theta_0\, b^{-j/(d/2-1)}, \qquad \theta_0 = \frac{\pi}{\ln H}. \tag{8}$$

The frequencies set only the positional resolution of the global layers. They do not change which keys are in the extrapolation regime, since that depends only on $\Delta_{\ln}$ (Eq. 3). Log-Ratio RoPE therefore remains length generalizable for any choice of b and θ_0.

Trainable Positional Resolution. The frequencies of Eq. 8 give every head the same frequencies. However, different heads specialize for different roles and require different positional resolution (Olsson et al., 2022; Wang et al., 2022). Each head h on a Log-Ratio layer learns one scalar $\lambda_h \in (0, \lambda_{\max})$ that scales its relation,

$$\Delta_h = \lambda_h(\ln m - \ln n) = \lambda_h \ln m - \lambda_h \ln n. \tag{9}$$

By Sec. 5.3, scaling the relation by λ_h scales the resolution bound by λ_h. A head with $\lambda_h > 1$ has higher positional resolution, and a head with $\lambda_h \to 0$ receives no rotation and reduces to NoPE. The relation is still a difference $f(m) - f(n)$ with $f = \lambda_h \ln(\cdot)$, so every head remains compute efficient (Prop. 1). The share of keys in the extrapolation regime does not depend on λ_h, since λ_h scales the trained range and every value of Δ by the same factor (Sec. F.2). We ablate the trainable scale, its initialization, and $\lambda_{\max}$ in Sec. F.4, Sec. F.5, and Sec. F.6.

7 Related Work

Prior work gives two accounts of RoPE's failure beyond the training length: the explicit positional encoding itself (Kazemnejad et al., 2023), or offsets outside the trained range (Chen et al., 2023; Men et al., 2024; Liu et al., 2024). Context-extension methods (Chen et al., 2023; Peng et al., 2026) act on it only up to a chosen target length (Lu et al., 2024), while Log-Ratio RoPE keeps almost all keys of the global layers in the trained range at every length (Sec. 6). The RoPE–NoPE hybrid (Yang et al., 2025a) was first motivated by the attention patterns of trained layers, which show where a model attends but not why. We instead compare hybrids that differ only in the encoding of the global layers (Sec. 4). A full discussion of related work is in Appendix G.

8 Experiments

We evaluate the Log-Ratio hybrid of Sec. 6 at 0.5B parameters. The models follow Qwen2-0.5B (Yang et al., 2024) and are trained from scratch at $S = 8$K on 10B FineWeb-Edu tokens (Lozhkov et al., 2024), about 20 tokens per parameter (Hoffmann et al., 2022). We train RoPE and three 3:1 hybrids whose sliding-window layers use RoPE and whose global layers use RoPE, NoPE, or Log-Ratio RoPE. We name each hybrid by its global layers, so the RoPE, NoPE, and Log-Ratio hybrids are the RoPE–RoPE, RoPE–NoPE, and RoPE–Log-Ratio hybrids in the local–global naming of Sec. 4. As context-extension baselines, we post-train RoPE and the RoPE hybrid with YaRN (Peng et al., 2026) for 0.5B tokens at 32K, applying YaRN only to the global layers of the hybrid. Only the YaRN models receive additional post-training. Appendix A gives the full training details. We run all evaluations with the LM Eval Harness (Gao et al., 2024).

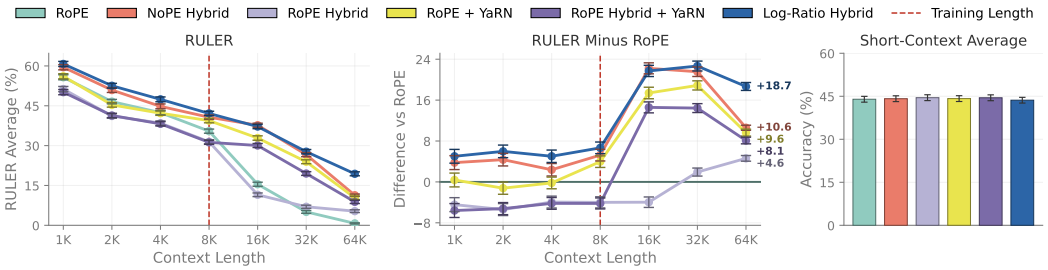

Figure 4: RULER and short-context accuracy. **Left:** RULER accuracy averaged over 13 tasks. **Middle:** Each model minus RoPE. **Right:** Zero-shot accuracy averaged over six short-context benchmarks. The red dashed line marks the training length. Error bars are 95% confidence intervals.

Short-Context Benchmarks. We measure zero-shot accuracy on HellaSwag (Zellers et al., 2019), ARC-Challenge, ARC-Easy (Clark et al., 2018), PIQA (Bisk et al., 2019), WinoGrande (Sakaguchi et al., 2019), and OpenBookQA (Mihaylov et al., 2018), whose inputs lie well inside the training length. All six models perform similarly, and none of the models we test shows a regression on short-context benchmarks (Fig. 4, right).

Long-Context Benchmarks. We evaluate on RULER (Hsieh et al., 2024) and report accuracy averaged over its 13 tasks at lengths from 1K to 64K (Fig. 4, left and middle). The Log-Ratio hybrid has the highest accuracy both within and beyond the training length, except at 16K, where its deficit lies within the 95% confidence interval. Within the training length, the differences between models are small compared with those beyond it. Beyond it, RoPE and the RoPE hybrid, whose global layers attend at out-of-distribution offsets, drop sharply (Sec. 3, Sec. 4). YaRN post-training improves both models, but the NoPE and Log-Ratio hybrids outperform both YaRN models without it. The effect of the inductive bias is clearest at 64K, eight times the training length, where the Log-Ratio hybrid scores 18.7 points above RoPE and 8.1 points above the NoPE hybrid.

This gap reflects a difference in design: Log-Ratio RoPE is designed for length generalization, whereas the NoPE hybrid relies on training to provide it. The NoPE hybrid outperforms the RoPE hybrid because RoPE fails at out-of-distribution offsets, not because NoPE length generalizes (Sec. 4). The Log-Ratio hybrid instead splits the trade-off of Theorem 1 across its local and global layers (Sec. 6).

9 Conclusion

We trained models from scratch that differ only in architecture to find where each architectural choice limits long-context performance. Removing positional encoding does not improve extrapolation, and within the training length RoPE learns in-context rules that NoPE does not at our scale (Sec. 2). RoPE fails beyond the training length mainly because of keys at offsets never seen in training, and a hybrid does not remove this failure because its global attention still reads far keys at these offsets (Sec. 3, Sec. 4). At these offsets RoPE is worse for retrieval than no positional encoding, which explains why the RoPE–NoPE hybrid outperforms the RoPE–RoPE hybrid. No relative positional encoding can keep almost all keys in the trained range while also keeping its positional resolution (Theorem 1). A hybrid with RoPE in the local attention and Log-Ratio RoPE in the global attention splits the two properties across layers. At 0.5B parameters, it matches RoPE on short-context benchmarks and improves RULER accuracy at 64K, eight times the training length, by 18.7 points over RoPE and 8.1 points over the RoPE–NoPE hybrid (Sec. 8). Without any post-training, it also outperforms the RoPE and RoPE–RoPE models post-trained with YaRN. RoPE helps at trained offsets but hurts at unseen ones, so it should be restricted to layers whose offsets stay within the trained range, such as local attention. We discuss the limitations of our study in Appendix H.

AI USE STATEMENT

In this work, we used generative AI tools for [implement methods]. We have not used generative AI tools for desiging experiments, and [Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results] are not applicable to this work. Additionally, we used generative AI tools for []. We have reviewed all AI-assisted work. [Elaborate. For example, "we cheked LLM-generated code where relevant"]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ekin Akyürek, Bailin Wang, Yoon Kim, and Jacob Andreas. In-context language learning: Architectures and algorithms. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=3Z9CRr5srL.

Zeyuan Allen-Zhu. Physics of language models: Part 4.1, architecture design and the magic of canon layers, 2026. URL https://arxiv.org/abs/2512.17351.

Noah Amsel, Gilad Yehudai, and Joan Bruna. Quality over quantity in attention layers: When adding more heads hurts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=y9Xp9NozPR.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models, 2023. URL https://arxiv.org/abs/2312.04927.

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, 2024. URL https://arxiv.org/abs/2308.14508.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful?, 2025. URL https://arxiv.org/abs/2410.06205.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models, 2020. URL https://arxiv.org/abs/2002.07028.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Ta-Chung Chi, Ting-Han Fan, Peter J. Ramadge, and Alexander I. Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation, 2022. URL https://arxiv.org/abs/2205.09921.

Ta-Chung Chi, Ting-Han Fan, Li-Wei Chen, Alexander Rudnicky, and Peter Ramadge. Latent positional information is in the self-attention variance of transformer language models without positional embeddings. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1183–1193, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.102. URL https://aclanthology.org/2023.acl-short.102/.

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, 2018. URL https://arxiv.org/abs/1803.05457.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Ellen Gilsenan-McMahon, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukáš Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Sebastian Ruder, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Shang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command a: An enterprise-ready large language model, 2025. URL https://arxiv.org/abs/2504.00698.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019. URL https://arxiv.org/abs/1901.02860.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng

Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Benjamin L. Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. The evolution of statistical induction heads: In-context learning markov chains, 2024. URL https://arxiv.org/abs/2402.11004.

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling?, 2025. URL https://arxiv.org/abs/2410.23771.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan

Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov,

Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information, 2022. URL https://arxiv.org/abs/2203.16634.

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers, 2020. URL https://arxiv.org/abs/2010.04245.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma, July 2025. URL https://trychroma.com/research/context-rot.

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?, 2024. URL https://arxiv.org/abs/2404.06654.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,

Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024. URL https://arxiv.org/abs/2310.06770.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers, 2023. URL https://arxiv.org/abs/2305.19466.

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers, 2024. URL https://arxiv.org/abs/2310.04418.

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, 2023. URL https://arxiv.org/abs/2307.03172.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation, 2024. URL https://arxiv.org/abs/2310.05209.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

Yi Lu, Jing Nathan Yan, Songlin Yang, Justin T. Chiu, Siyu Ren, Fei Yuan, Wenting Zhao, Zhiyong Wu, and Alexander M. Rush. A controlled study on long context extension and generalization in llms, 2024. URL https://arxiv.org/abs/2409.12181.

Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. Base of rope bounds context length, 2024. URL https://arxiv.org/abs/2405.14591.

Meta Superintelligence Lab. Muse-Glimmer-30B. Hugging Face model repository, August 2026. URL https://huggingface.co/meta-models/Muse-Glimmer-30B. Accessed: 2026-09-26.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018. URL https://arxiv.org/abs/1809.02789.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Bohan Zhang, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wenting Zhan, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2026. URL https://arxiv.org/abs/2412.16720.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2026. URL https://arxiv.org/abs/2309.00071.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2022. URL https://arxiv.org/abs/2108.12409.

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, 2023. URL https://arxiv.org/abs/1910.10683.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

Abulhair Saparov, Srushti Pawar, Shreyas Pimpalgaonkar, Nitish Joshi, Richard Yuanzhe Pang, Vishakh Padmakumar, Seyed Mehran Kazemi, Najoung Kim, and He He. Transformers struggle to learn to search, 2025. URL https://arxiv.org/abs/2412.04703.

Connor Schenck, Isaac Reid, Mithun George Jacob, Alex Bewley, Joshua Ainslie, David Rendleman, Deepali Jain, Mohit Sharma, Avinava Dubey, Ayzaan Wahid, Sumeet Singh, René Wagner, Tianli Ding, Chuyuan Fu, Arunkumar Byravan, Jake Varley, Alexey Gritsenko, Matthias Minderer, Dmitry Kalashnikov, Jonathan Tompson, Vikas Sindhwani, and Krzysztof Choromanski. Learning the ropes: Better 2d and 3d position encodings with string, 2025. URL https://arxiv.org/abs/2502.02562.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations, 2018. URL https://arxiv.org/abs/1803.02155.

Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer, 2022. URL https://arxiv.org/abs/2212.10554.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy,

Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv.org/abs/2211.00593.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Bowen Yang, Bharat Venkitesh, Dwarak Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr Locatelli. Rope to nope and back again: A new hybrid attention strategy, 2025a. URL https://arxiv.org/abs/2501.18795.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule, 2025b. URL https://arxiv.org/abs/2412.06464.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019. URL https://arxiv.org/abs/1910.07467.

Yu Zhao, Yuanbin Qu, Konrad Staniszewski, Szymon Tworkowski, Wei Liu, Piotr Miłoś, Yuxiang Wu, and Pasquale Minervini. Analysing the impact of sequence composition on language model pre-training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7897–7912. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.427. URL http://dx.doi.org/10.18653/v1/2024.acl-long.427.

A Model training details

The 64M small models use pre-norm RMSNorm (Zhang & Sennrich, 2019), SwiGLU (Shazeer, 2020), query-key normalization (Henry et al., 2020), and the Llama-2 tokenizer (Touvron et al., 2023). We train each model at three learning rates on 5B FineWeb-Edu tokens (Lozhkov et al., 2024), roughly four times the compute-optimal budget for this model size (Hoffmann et al., 2022).

The 0.5B models follow Qwen2-0.5B (Yang et al., 2024), except that we replace grouped-query attention with multi-head attention and apply the same query-key normalization used in the small models. We train them from scratch on 10B FineWeb-Edu tokens, or about 20 tokens per parameter. RoPE + YaRN is post-trained the RoPE model for 0.5B tokens at length 32K with a scaling factor of four (Peng et al., 2026).

Across both scales, standard RoPE uses base $b = 10^6$. The hybrid models use a 512-token sliding window in three out of every four layers and full attention in the remaining layers. The Log-Ratio model in the main comparison uses an unwrapped frequency ladder designed for 128K (Sec. 6), with a learned multiplier $\gamma_h = 2\,\sigma(a_h)$ for each global head and $a_h \sim \mathcal{N}(0, 1)$ at initialization. All models use intra-document attention masking to prevent attention across packed document boundaries (Zhao et al., 2024).

Table 1 lists the architecture and optimization settings. We use AdamW for training all models (Loshchilov & Hutter, 2019). For the small-model comparisons, we select the learning rate with the lowest test loss on FineWeb-Edu loss at lengths up to the training length.

Table 1: Training hyperparameters for the models trained from scratch. The sequence batch includes all devices and gradient-accumulation steps.

	Small	Large
Layers	12	24
Width	512	896
Intermediate width	1024	4864
Heads	4	8
Head dimension	128	112
Vocabulary	32K	152K
Training length S	8K	8K
Tokens per step	262K	262K
Sequences per step	32	32
Training tokens	5B	10B
Optimizer	AdamW	AdamW
Weight decay	0.01	0.01
Gradient clipping	1.0	1.0
Peak learning rate	$5e{-4}, 8e{-4}, 2e{-3}$	$5e{-4}$
Warmup steps	1,000	1,000
Decay ratio	10%	10%
Decay shape	Linear	Cosine

B Evaluation Metrics

B.1 Long-document loss

The loss curves report average next-token negative log-likelihood on 500 long documents from the Dolma 3 Longmino Mix (Olmo et al., 2025), the long-context pretraining data of Olmo 3. At each evaluation length L, we truncate the same 500 documents to their first L tokens and average the per-token loss across all positions. Error bars are 95% confidence intervals over the 500 documents.

B.2 KEY-VALUE RECALL

Reasoning over a long context requires finding information stated earlier and using it. Key-value recall tests the simplest form of this. Its rule is the one implemented by induction heads, which may underlie most in-context learning (Olsson et al., 2022), and heads that copy from the context are needed for chain-of-thought reasoning, where the model refers back to the question and its own earlier steps (Allen-Zhu, 2026; Saparov et al., 2025).

Key-value recall is a multi-query needle-in-a-haystack task in the style of RULER (Hsieh et al., 2024), with content-free keys and values as in the associative-recall tasks of Arora et al. (2023). The haystack is one of the same 500 Dolma 3 documents used in Sec. B.1 truncated to the evaluation length L. Into it we plant $K = 8$ records of the form "$\backslash n\ a\ b\ v$" at evenly spaced depths, where the two-token key (a, b) and the value v are random token IDs. The records carry no meaning, so the only way to answer is to find the earlier occurrence of the key and copy the token that followed it. We do not put a separator between key and value: with one, the token immediately preceding every value is identical across records, and a single-token induction match lands on all of them at once. Keys are unique within a haystack, and values are drawn without replacement, so a wrong retrieval can never be scored correct by accident. Because each key is queried on its own, a single query is RULER's single-needle setting with the remaining seven records as distractors.

For example, with $K = 3$ records, the sequence and one of its queries are

$$\text{[Context]}\ \underbrace{A\ B\ 3}_{\text{Key-Value}}\ \text{[Context]}\ \underbrace{C\ D\ 7}_{\text{Key-Value}}\ \text{[Context]}\ \underbrace{E\ F\ 1}_{\text{Key-Value}}\ \text{[Context]}\ \rightarrow\ \underbrace{C\ D\ ?}_{\text{Query}}$$

where each [Context] is a span of the Dolma 3 document, $K = 3$ records are planted, and the correct output for the query is 7. The letters and digits stand in for random token IDs, and the newline token that precedes each record and the query is omitted. We hold K fixed and grow only the haystack, which isolates retrieval at distance. Filling the whole context with records instead, as in multi-query associative recall (Arora et al., 2023), makes capacity rather than distance the binding constraint. At 16K, that is thousands of competing keys, and every model is at chance.

To score a record, we append "$\backslash n\ a\ b$" to the context and read the next-token distribution, restricted to a pool of 64 candidate values. The query is correct if the argmax over the pool is the planted value, so the chance is $1/64$. Nothing is generated, and no instruction is given, which suits base models at this scale (64M). Each plotted point is the accuracy over $500 \times 8 = 4000$ queries: 500 haystacks, each with $K = 8$ planted records, every record queried once. Since the 8 queries of one haystack share a context, they are not independent, so we first average the 8 queries within each haystack and then report 95% confidence intervals over the 500 haystack means rather than over the 4000 queries.

B.3 GRAPH WALK

Graph walk is a fully synthetic in-context learning task. The whole context is made of walks on one hidden graph, so the context does not contain natural language. We draw a hidden successor graph over 64 states, where each state has exactly one successor. The graph is a random permutation whose cycles all have length at least 16, so a walk of 8 states never visits the same state twice. Each state is shown as a single token, drawn at random from a fixed set of 512 word tokens in the vocabulary, with a new assignment for every graph. The context is a sequence of walk lines $[\backslash\mathrm{n}, n_0, n_1, n_2, n_3, n_4, n_5, n_6, n_7]$, each starting at a random state and following the graph for 8 states.

We make the task harder by letting several states share one token. With V tokens, each token labels $64/V$ states, so an earlier occurrence of the current token may have been followed by a different token. We choose the labels so that the last c tokens of a walk always identify the current state, and so its successor. Within a walk line, the token sequence is therefore a deterministic Markov chain of order c (Edelman et al., 2024). We use three settings: $(V, c) = (64, 1), (16, 2), (8, 3)$. With $V = 64$, every state has its own token.

Fig. 5 shows a small graph for each setting and one example context. With $c = 2$, the token B labels two states: one that follows C and leads to D, and one that follows A and leads to A. In the query, the most recent B was followed by A, so copying the token after the last B gives the wrong answer.

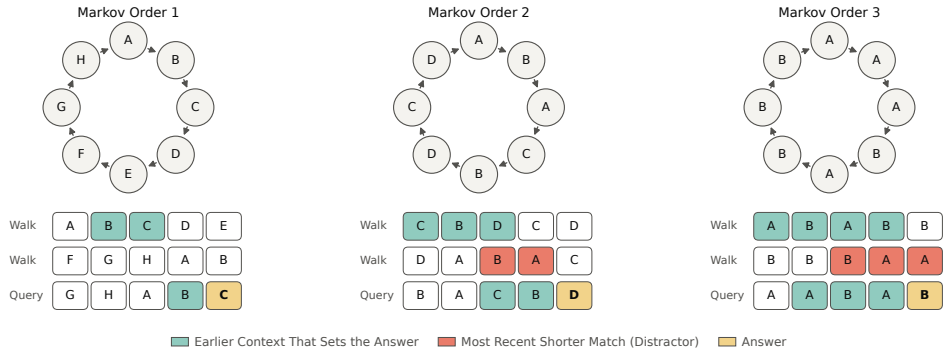

Figure 5: Graph walk at Markov order 1, 2 and 3, where the last c tokens set the answer. Each ring is a hidden successor graph with 8 states, and each state shows its token. The rows below are two walks from the context and a query walk. Green marks the earlier context that sets the answer, red marks the most recent shorter match, which points to a wrong answer, and yellow marks the answer. For readability, each ring has 8 states and each walk shows 5 tokens. The actual task uses a graph with 64 states, and each walk line visits 8 consecutive states.

The model has to find the earlier occurrence of the pair $C\,B$ and copy the token that followed it. With $c = 3$, the same holds one level up: the most recent pair $B\,A$ was followed by A, and only the triple $A\,B\,A$ gives the answer B.

We score every step of every walk, n_1 through n_7, in a single forward pass over an 8K-token stream. At each step we read the next-token distribution restricted to the V tokens, and the step is correct if the argmax is the true successor. Because attention is causal, a prediction does not depend on later tokens, so cutting the stream at an evaluation length L leaves the predictions before L unchanged. One pass therefore gives the accuracy at every evaluation length at once. At evaluation length L, we report the accuracy over the steps in the second half of the context, from $L/2$ to L. We keep only steps where the last c tokens lie inside the current walk line and the model has already seen the same state transition earlier in the context. Other steps cannot be predicted from the context. The chance level is $1/V$.

As the walk lines repeat, every state recurs about every 72 tokens. The evidence for a correct answer is therefore always close by, and a drop in accuracy with evaluation length means the model breaks at that length, not that the evidence moved out of reach. Each stream has 6,370 scored steps before filtering, and we evaluate 8 streams for each setting, each with its own graph. The steps of one stream all come from the same 64 states, so they are not independent. We therefore report 95% confidence intervals over the 8 per-graph accuracies rather than over the steps.

C RoPE ANALYSIS

C.1 KEY-VALUE RECALL WITH ONE MOVED RECORD

Fig. 2 plants a single record and moves it through the context. We split the context into 1K blocks and place the record at the center of one block per run. Each document has one key and one value, and they stay the same across blocks and models, so only the position of the record changes. We use the same 500 documents and score as in Sec. B.2, and pool the blocks into the two halves of the context. To check that the result does not depend on this layout, we also run the standard task, with eight records at fixed positions. Recall for the model trained at 8K is the same in both layouts: 14% on the first half and 48% on the second half at 16K with one record, against 17% and 48% with eight records.

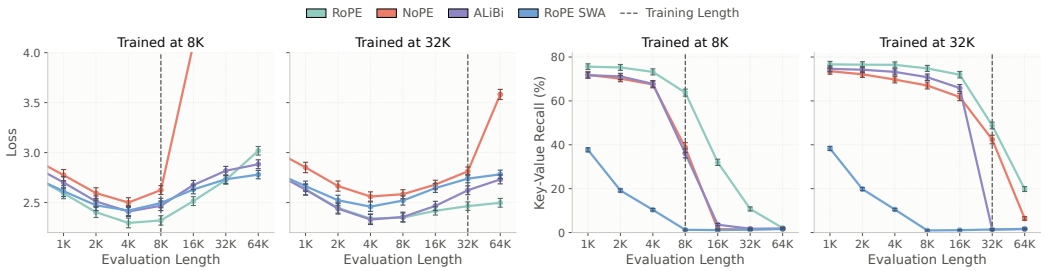

Figure 6: RoPE, NoPE, SWA and ALiBi at $H = 4$. Dashed lines mark the training length. **Left two panels:** Loss (lower is better). **Right two panels:** Key-value recall (higher is better). Error bars are 95% confidence intervals.

D Length-Generalization Problem Extended

D.1 Inductive biases and context utilization

We compare RoPE, NoPE, ALiBi (Press et al., 2022), and sliding-window attention (SWA) (Beltagy et al., 2020; Jiang et al., 2023), while keeping the remaining training setup fixed (Appendix A). All four are trained at $S \in \{8\text{K}, 32\text{K}\}$. Our SWA model uses RoPE for positional information but restricts each layer to attending within a fixed local window. In contrast, ALiBi provides explicit relative positional information via a position-dependent bias added directly to the attention logits.

As shown in Fig. 6, these architectural choices lead to substantially different context utilization, especially at $S = 32\text{K}$. SWA performs poorly on key–value recall (Sec. B.2) even within its theoretical receptive field (Beltagy et al., 2020; Jiang et al., 2023), showing that reachability in theory can substantially overestimate usable context. Training SWA at 32K instead of 8K does not change its recall: in both models recall is below 40% at 1K and near zero from 8K.

ALiBi exhibits a similar limitation despite allowing global attention. When trained at 8K, ALiBi achieves recall comparable to NoPE within the training context, but when trained at 32K it fails to use the full 32K window reliably and performs substantially worse than both NoPE and RoPE. In contrast, NoPE learns to use distant information despite receiving no explicit positional information, while RoPE achieves the strongest context utilization across these settings. Thus, supplying explicit positional information is not necessarily sufficient to improve context utilization: the particular architectural form through which that information is introduced can instead limit how effectively the model learns to use long-range context.

These limitations come from the inductive bias of each model. SWA at both training lengths and ALiBi trained at 32K have lower loss than NoPE inside their training context, so they model nearby tokens well. Yet they have lower recall. Each SWA layer attends only within its window, and the ALiBi bias lowers the logit of a key in proportion to its distance. Both constraints limit what the model can learn from distant tokens. An inductive bias should therefore be designed so that its constraints on long-range attention are not too strong.

D.2 Training length and head dim with RoPE

As context length grows, attention must distinguish relevant information among more tokens using a fixed-dimensional query–key representation. Prior work establishes worst-case expressivity limits that depend on attention head dimension (Bhojanapalli et al., 2020; Amsel et al., 2025), but does not determine the head dimension required in practice for long-context models. If representational capacity is the primary bottleneck, increasing the head dimension d should substantially improve long-context performance.

To examine the representational limitation imposed by the attention head dimension, we vary the training length $S \in \{1\text{K}, 8\text{K}, 16\text{K}, 32\text{K}\}$ and head dimension $d \in \{32, 64, 128, 256\}$ in a small model with RoPE. We use RoPE because it has the strongest context utilization among the encodings in Fig. 6, including at $S = 32\text{K}$, and because it is the positional encoding of widely used large

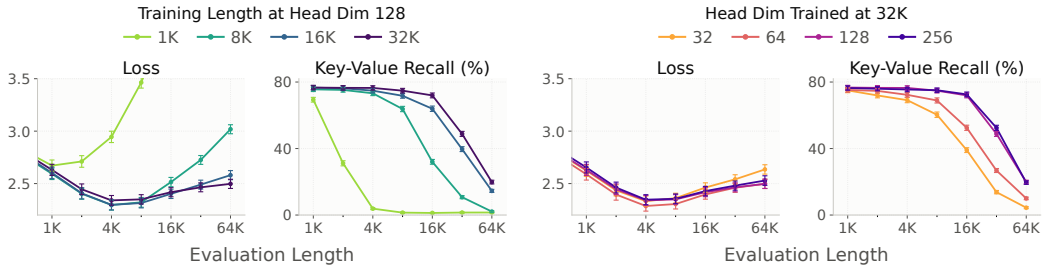

Figure 7: Training at longer context lengths extends the range of low loss and useful recall, and head dim mainly affects recall. **Left:** RoPE models with head dim 128, colours indicate training length. **Right:** RoPE models trained at 32K, colours indicate head dim. Loss is next-token loss (lower is better). Error bars show 95% confidence intervals. Fig. 8 shows every pair of head dim and training length.

models (Grattafiori et al., 2024; DeepSeek-AI et al., 2025; Yang et al., 2024). If representational capacity is the primary limitation, increasing d should extend the range over which the model can retrieve distant information. If degradation persists beyond the training length once d is sufficiently large, it instead indicates a distinct limitation in length generalization.

Fig. 7 shows that training on longer contexts improves context utilization by extending the range over which RoPE models maintain low loss and useful key–value recall. Head dimension also matters for context utilization: models with $d = 32$ and $d = 64$ exhibit substantially worse recall, whereas $d = 128$ and $d = 256$ perform similarly within the training context. However, increasing d does not resolve length generalization: all models degrade sharply once evaluation exceeds the training length, with little additional benefit from increasing d beyond 128. Even when trained at 32K, the RoPE model can effectively utilize context throughout its training window but fails rapidly beyond it. Thus, sufficient head dimension is important for context utilization, but representational capacity alone does not explain the failure to generalize to longer contexts. Training on longer sequences is costly and limited by the supply of long documents, so Sec. 3 instead asks which part of the input is out of distribution beyond the training length.

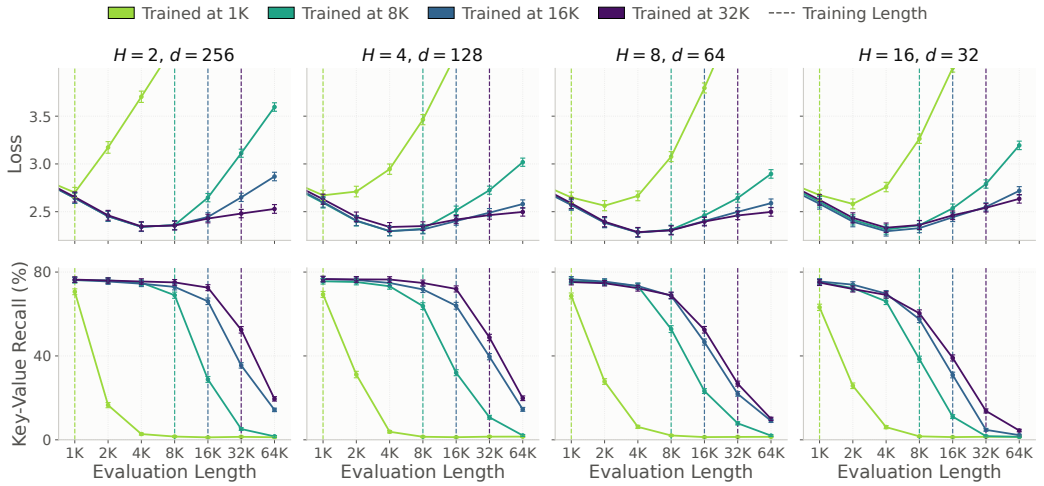

Figure 8: Training at longer context lengths generally extends the range of low loss and useful recall. Columns vary head count and colours indicate training context length. **Top:** Next-token loss (lower is better). **Bottom:** Key-value recall (higher is better). Error bars show 95% confidence intervals.

Fig. 8 shows every pair of head count $H \in \{2, 4, 8, 16\}$ and training length for the small RoPE model. The width is fixed at 512, so the head dim is $512/H$. Fig. 7 shows a subsection of this

figure. We see the same trend where RoPE's performance degrades after training length. But for every training length, RoPE's window performance is better.

E Ideal Positional Encoding

E.1 Form of the kernel

Under Def. 1, condition (i) makes every $\mathrm{R}(\Delta)$ orthogonal. Taking $\Delta = \delta = 0$ in condition (ii) gives $\mathrm{R}(0) = \mathrm{R}(0)^2$, and since $\mathrm{R}(0)$ is invertible, $\mathrm{R}(0) = \mathrm{I}$. Condition (ii) with continuity makes R a continuous one-parameter subgroup of the orthogonal group. Such a subgroup has the form $\mathrm{R}(\Delta) = \exp(\Delta \mathrm{A})$ for a fixed skew-symmetric generator $\mathrm{A} \in \mathbb{R}^{d\times d}$ (Schenck et al., 2025, Theorem 3.2). Schenck et al. (2025) obtain the group law from $\Delta = m - n$. Here condition (ii) states it directly in Δ, so the same form holds for every relation $\Delta(m, n)$, not only for $m - n$. A real orthogonal change of basis puts A into 2×2 blocks $\theta_j \mathrm{J}$, with J the rotation generator, together with a subspace on which it vanishes. Inside block j, $\exp(\Delta \mathrm{A})$ is a plane rotation by $\theta_j \Delta$. This form is smooth in Δ, which is what the interpolation argument in Sec. E.4 relies on.

E.2 Compute efficiency

An encoding is compute efficient (Sec. 5.1) when the query and the key are transformed by maps of their own positions, $q_m = \mathrm{P}(m)\, q$ and $k_n = \mathrm{P}(n)\, k$. The logits are then a single product $QK^\top$, so training can use fused attention kernels and decoding can cache the keys. If the encoding of a key depends on the query position, every query needs its own copy of every key, which takes $O(L^2 d)$ memory instead of $O(Ld)$. By Sec. E.1, the compute efficiency condition reads

$$\mathrm{P}(n)^\top \mathrm{P}(m) \;=\; \mathrm{R}(\Delta(m, n)) \;=\; \exp\big(\Delta(m, n)\, \mathrm{A}\big). \tag{10}$$

Requiring compute efficiency restricts the relation.

Proposition 1 (Compute-efficient relations)**.** *Let* R *be one-to-one. An encoding that satisfies relativity is compute efficient if and only if*

$$\Delta(m, n) \;=\; f(m) - f(n) \tag{11}$$

for some position map $f : \mathbb{R} \to \mathbb{R}$*, and then one can take* $\mathrm{P}(m) = \exp(f(m)\mathrm{A})$.

Proof of Prop. 1. If $\Delta(m, n) = f(m) - f(n)$, take $\mathrm{P}(m) = \exp(f(m)\mathrm{A})$. Then $\mathrm{P}(n)^\top \mathrm{P}(m) = \exp\big((f(m) - f(n))\mathrm{A}\big) = \mathrm{R}(\Delta(m, n))$.

Conversely, suppose Eq. 10 holds. Setting $n = m$ gives $\mathrm{P}(m)^\top \mathrm{P}(m) = \mathrm{R}(0) = \mathrm{I}$, so every $\mathrm{P}(m)$ is orthogonal. Fix a reference position p and let $f(m) = \Delta(m, p)$. Inserting $\mathrm{P}(p)\mathrm{P}(p)^\top = \mathrm{I}$ and using $\mathrm{R}(x)^\top = \mathrm{R}(-x)$,

$$\mathrm{R}(\Delta(m, n)) = \mathrm{P}(n)^\top \mathrm{P}(p)\, \mathrm{P}(p)^\top \mathrm{P}(m) = \mathrm{R}(f(n))^\top \mathrm{R}(f(m)) = \mathrm{R}\big(f(m) - f(n)\big).$$

Since R is one-to-one, $\Delta(m, n) = f(m) - f(n)$. □

$\mathrm{R}(x) = \mathrm{I}$ only if $\theta_j x \in 2\pi\mathbb{Z}$ for every j, so R is one-to-one whenever two frequencies of A have an irrational ratio, as the geometric frequencies of RoPE and Log-Ratio RoPE do.

The ratio $\Delta(m, n) = 1 - n/m$ is not a difference $f(m) - f(n)$. Every difference satisfies $\Delta(m, n) + \Delta(n, p) = \Delta(m, p)$ for $p \le n \le m$. For the ratio with $m = 4$, $n = 2$ and $p = 1$, the left side is $1/2 + 1/2 = 1$ and the right side is $3/4$. By Prop. 1, the ratio is therefore not compute efficient for any one-to-one R.

E.3 Derivation of Eq. 4

Let the keys at positions i and j share the content k. By Eq. 2, the logit gap between them is

$$\langle q_m,\, k_i\rangle - \langle q_m,\, k_j\rangle \;=\; \big\langle \big(\mathrm{R}(\Delta(m, i)) - \mathrm{R}(\Delta(m, j))\big) q,\; k\big\rangle. \tag{12}$$

By condition (ii) of Def. 1, $\mathrm{R}(\Delta(m,i)) = \mathrm{R}\big(\Delta(m,i) - \Delta(m,j)\big)\,\mathrm{R}(\Delta(m,j))$, so the gap is $\langle \mathrm{M}u,\, k\rangle$, where $\mathrm{M} = \mathrm{R}\big(\Delta(m,i) - \Delta(m,j)\big) - \mathrm{I}$ and $u = \mathrm{R}(\Delta(m,j))\,q$. Since $\mathrm{R}(\Delta(m,j))$ is orthogonal, u is a unit vector whenever q is, and every unit vector u is reached by some unit q. The largest value of $\langle \mathrm{M}u,\, k\rangle$ over unit u and k is the spectral norm $\|\mathrm{M}\| = \|\mathrm{I} - \mathrm{R}(\Delta(m,i) - \Delta(m,j))\|$. The factor $\mathrm{R}(\Delta(m,j))$ changes the gap for a fixed q but not its largest value over unit q, so the positional resolution of two keys depends only on the difference $\Delta(m,i) - \Delta(m,j)$ between their values of Δ. When this norm is zero, no query can tell the two positions apart.

E.4 Rate of the resolution norm

Here we bound how fast the norm in Eq. 4 falls as two keys take closer values of Δ. Let x be the difference between their values of Δ, so that the norm is $\|\mathrm{I} - \exp(x\mathrm{A})\|$. In the block basis of Sec. E.1, $\exp(x\mathrm{A})$ is a rotation by $\theta_j x$ inside block j. For a plane rotation by φ, both singular values of $\mathrm{I} - \mathrm{R}(\varphi)$ equal $2|\sin(\varphi/2)|$, and the spectral norm of a block-diagonal matrix is the largest of its blocks, so

$$\big\|\mathrm{I} - \exp(x\mathrm{A})\big\| = \max_j\, 2\big|\sin(\theta_j x/2)\big|.$$

Write $\theta_{\max} = \max_j |\theta_j|$ for the fastest frequency. Since $2|\sin(\varphi/2)| \le |\varphi|$, the norm is at most $\theta_{\max}|x|$. Since $2|\sin(\varphi/2)| \ge 2|\varphi|/\pi$ for $|\varphi| \le \pi$, it is at least $(2/\pi)\,\theta_{\max}|x|$ while $\theta_{\max}|x| \le \pi$. The two bounds differ by a fixed factor, so in that range the resolution falls at the rate of $|x|$ itself. Past that range the sine wraps and the norm oscillates instead of tracking $|x|$. When A has more than one frequency, one frequency completing a whole turn does not remove the resolution between far keys, since the norm is small only when every frequency is close to a whole turn at once.

The bound shows how the choice of Δ sets the resolution. When $\Delta(m,n) = m - n$, as in RoPE, two keys c positions apart differ in Δ by c wherever they sit, so their resolution does not change with position. When $\Delta(m,n) = 1 - n/m$, the ratio from Sec. 5.2, the same two keys differ by only c/m, so their resolution is at most $\theta_{\max}\, c/m$. The ratio keeps every key in $[0,1)$, so as the context grows the keys crowd into this interval. The property that makes the ratio length generalizable, keeping its keys in the trained range, is therefore what lowers its positional resolution.

The same bound shows that R changes slowly with Δ. Since $\mathrm{R}(\Delta) - \mathrm{R}(\Delta') = \mathrm{R}(\Delta')\big(\mathrm{R}(\Delta - \Delta') - \mathrm{I}\big)$ and $\mathrm{R}(\Delta')$ is orthogonal,

$$\big\|\mathrm{R}(\Delta) - \mathrm{R}(\Delta')\big\| = \big\|\mathrm{I} - \mathrm{R}(\Delta - \Delta')\big\| \le \theta_{\max}\,|\Delta - \Delta'|.$$

A value of Δ inside the trained range therefore gets a kernel within $\theta_{\max}$ times its distance to the nearest value in Δ_S (Sec. 5.2). This is why a key in the interpolation regime counts as in distribution even when its value of Δ never occurs in training: its kernel is close to one the model has seen. This requires Δ_S to have no large gaps inside the trained range. It holds for the relations in this paper. For the difference, Δ_S contains every integer in the trained range. For the ratio and the log ratio, neighboring values in Δ_S are at most about $1/S$ apart.

E.5 Proof of Theorem 1

Fix a training length S and a query at position m, and let $D(S) = \max \Delta_S - \min \Delta_S$ be the width of the trained range. By Eq. 3, $K = (1 - \phi_S(m))\, m$ keys of the query have values of Δ inside the trained range. Sort these values as $v_1 \le \cdots \le v_K$. The $K - 1$ gaps $v_{k+1} - v_k$ add up to at most $D(S)$, so the smallest gap is at most $D(S)/(K-1)$. Let i and j be two keys with this gap. By Def. 3 and the bound $\|\mathrm{I} - \mathrm{R}(x)\| \le \theta_{\max}|x|$ of Sec. E.4, whenever $K \ge 2$,

$$r_m \;\le\; \rho_m(i,j) \;\le\; \theta_{\max}\big|\Delta(m,i) - \Delta(m,j)\big| \;\le\; \frac{\theta_{\max}\, D(S)}{(1 - \phi_S(m))\, m - 1}. \tag{13}$$

The argument does not use the order of the keys, so i and j can be neighbors or far apart.

If R is length generalizable, there is an S with $\sup_m \phi_S(m) \le 1/2$. For this S and every $m \ge 4$, Eq. 13 gives $r_m \le \theta_{\max} D(S)/(m/2 - 1)$, so $r_m = O(1/m)$. The second claim of Theorem 1 is the contrapositive: if $m\, r_m$ is unbounded, then r_m is not $O(1/m)$, so R is not length generalizable.

The theorem is a statement about the rate at which r_m falls. Every encoding that satisfies relativity has $r_m \to 0$ as m grows, whether or not it is length generalizable. The kernels $\mathrm{R}(\Delta(m,n))$ of the

m keys lie in the closure of $\{\exp(x\mathrm{A}) : x \in \mathbb{R}\}$, which is compact. For any $\varepsilon > 0$, once m exceeds the number of balls of radius $\varepsilon/2$ needed to cover this set, two of the kernels lie in the same ball. Since $\|\mathrm{R}(\Delta(m,i)) - \mathrm{R}(\Delta(m,j))\| = \rho_m(i,j)$ (Sec. E.4), this gives $r_m < \varepsilon$. When A has k distinct frequencies, this closure is a torus of dimension at most k, and the covering argument alone gives a rate of order $m^{-1/k}$. A length generalizable encoding is forced to the faster rate $1/m$, since almost all of its keys lie in an interval of fixed width.

The bound in Eq. 13 decays only when the number of keys in the trained range grows with m. RoPE keeps only S keys in the trained range, so its bound stays constant, while the ratio keeps almost all keys there, so its bound falls like $1/m$. The bound is a worst case: it holds for keys with identical content, whose positions reach attention only through R.

F Log-Ratio RoPE Extended

In this section, we derive the highest frequency θ_0 of Eq. 8 and ablate different hyperparameters in the log-ratio RoPE to examine their effects. In each subsection, we train two models from scratch, varying only a given hyperparameter. Generally, having trainable resolution parameters with $\lambda_{\max} = 2$ and random initialization helps the model perform well in-distribution without incurring additional extrapolation cost.

F.1 Highest frequency

We choose θ_0 from the positional resolution of Def. 3. The highest frequency of Eq. 8 is $\theta_{\max} = \theta_0$. For a query at position m, every key has $\Delta_{\ln}(m,n) = \ln(m/n) \in [0, \ln m]$, so two of its keys i and j differ in $\Delta_{\ln}$ by some x with $|x| \le \ln m$. By Sec. E.4,

$$\tfrac{2}{\pi}\,\theta_0|x| \;\le\; \rho_m(i,j) \;\le\; \theta_0|x| \qquad \text{whenever } \theta_0|x| \le \pi.$$

In this range, every term $2|\sin(\theta_j x/2)|$ of the resolution grows with $|x|$. Two keys that are farther apart in $\Delta_{\ln}$ are therefore never harder to separate than two closer keys, and the resolution of every pair is set by its gap in $\Delta_{\ln}$ up to a fixed factor. Outside this range, the term of the highest frequency decreases as $|x|$ grows, and the resolution no longer follows the gap.

We want this range to cover every pair of keys of every query up to a horizon H. The largest gap of such a query is $\ln H$, so this requires $\theta_0 \le \pi/\ln H$. Both bounds above grow with θ_0, so a larger θ_0 raises the resolution of every pair. $\theta_0 = \pi/\ln H$ is therefore the largest value for which the resolution grows with the gap for every query up to position H. A larger θ_0 would raise the resolution of close keys, but far pairs of queries near H would leave this range. A smaller θ_0 keeps every pair in range but lowers the resolution of every pair. Past H, some pairs leave this range, but their resolution does not vanish, since the lower frequencies still separate them (Sec. E.4). The choice of θ_0 does not affect length generalization, since the keys in the extrapolation regime depend only on $\Delta_{\ln}$ (Eq. 3). We use $H = 128$K.

F.2 Per-head scale

Each head h scales every angle by λ_h (Eq. 9), so the phase of pair j in Eq. 7 becomes

$$\alpha_{hj} = \lambda_h\,\theta_j \ln\frac{m}{n}. \tag{14}$$

With $\theta_0 = \pi/\ln H$, the fastest pair of head h turns half a turn at position H^{1/λ_h}. A head with $\lambda_h > 1$ resolves nearby positions more finely, but its fastest pair wraps before H. A head with $\lambda_h < 1$ has lower resolution and does not wrap until after H. In both cases the share of keys in the extrapolation regime is unchanged. The trained range of head h is $[0, \lambda_h \ln S]$, and $\lambda_h \ln(m/n)$ lies in it exactly when $\ln(m/n) \le \ln S$, which is the condition for $\lambda_h = 1$.

F.3 Bases

In this ablation, we train three 0.5B models on different frequency bases, 2, 32 and 256, without a trainable resolution parameter. The slowest pair turns b times slower than the fastest, so only half as

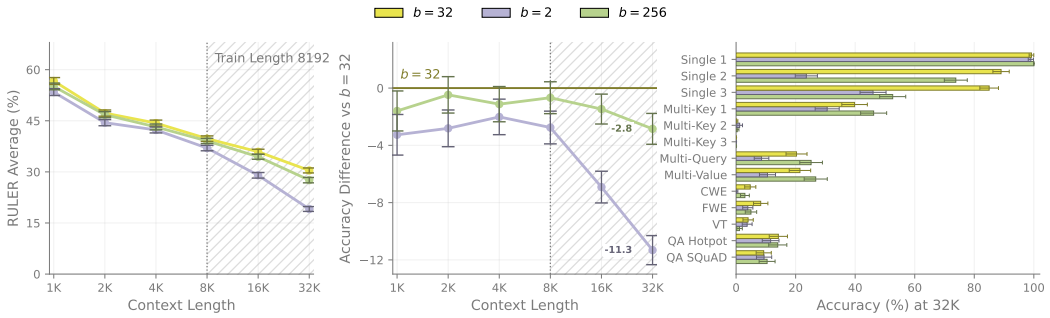

Figure 9: Frequency bases $b = 2$ and $b = 256$ against $b = 32$. RULER from 1K to 32K, panels as in Fig. 10 with the $b = 32$ frequency as the zero line, per-task accuracy at 32K.

fast at $b = 2$. Fig. 9 shows that $b = 32$ beats $b = 2$ both inside the training window and outside, by 11.3 points at 32K. $b = 256$ stays within 2 points of $b = 32$ up to 16K and falls 2.8 points behind at 32K.

F.4 Trainable Resolution Parameter

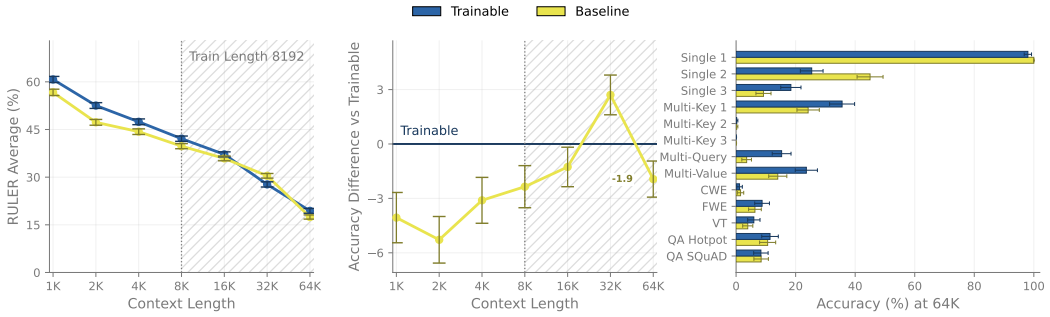

Figure 10: Trained speed against a fixed speed of $\lambda_h = 1$ on every head, both with the base $b = 32$ ladder. RULER (13 tasks, task-averaged) from 1K to 64K, higher is better. **Left:** Accuracy against evaluation length, the shaded region being past the training length. **Middle:** The variant minus the main-comparison model, with the gap at the last length labeled. **Right:** Per-task accuracy at 64K. Error bars are 95% confidence intervals.

In this ablation, we train two 0.5B models with and without a trainable resolution parameter, both with $b = 32$. Here, the trainable resolution parameter has $\lambda_{\max} = 2$ and is initialized with $\mathcal{N}(0, 1)$ across different heads for a given global layer.Fig. 10 shows that allowing the model to tune its own frequency resolution helps the model better resolve different positions within its training context window. Importantly, including this resolution parameter does not hurt OOD performance. Even when comparing performance at 64K, having the trainable parameters helps performance.

F.5 Trainable Resolution Parameter Init

In this ablation, we train two 0.5B models with the same trainable resolution parameter, $\lambda_{\max} = 2$, and change only the mean and variance of its initialization. The first model draws $a_h \sim \mathcal{N}(0, 1)$, so λ_h starts spread out across heads. The second draws $\lambda_h \sim \mathcal{N}(1, 0.1)$ and sets $a_h = \sigma^{-1}(\lambda_h/\lambda_{\max})$, so every head starts close to $\lambda_h = 1$. Fig. 11 shows that $a_h \sim \mathcal{N}(0, 1)$ leads within the training window, and $\lambda_h \sim \mathcal{N}(1, 0.1)$ closes the gap at 16K and leads at 32K. However, the effect of this is not as strong as in Fig. 9.

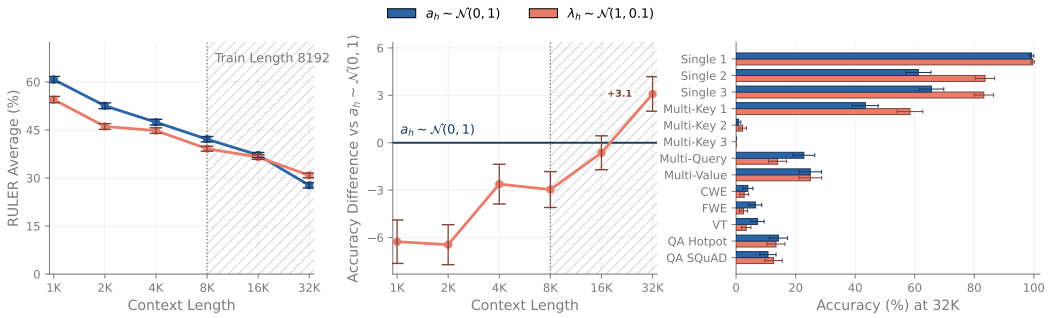

Figure 11: Initialization of the trained speed, $a_h \sim \mathcal{N}(0,1)$ against $\lambda_h \sim \mathcal{N}(1,0.1)$, both with $\lambda_{\max} = 2$. RULER (13 tasks, task-averaged) from 1K to 32K, higher is better. **Left:** Accuracy against evaluation length, the shaded region being past the training length. **Middle:** The variant minus the main-comparison model, with the gap at 32K labelled. **Right:** Per-task accuracy at 32K. Error bars are 95% confidence intervals.

F.6 Trainable Resolution Parameter $\lambda_{\max}$

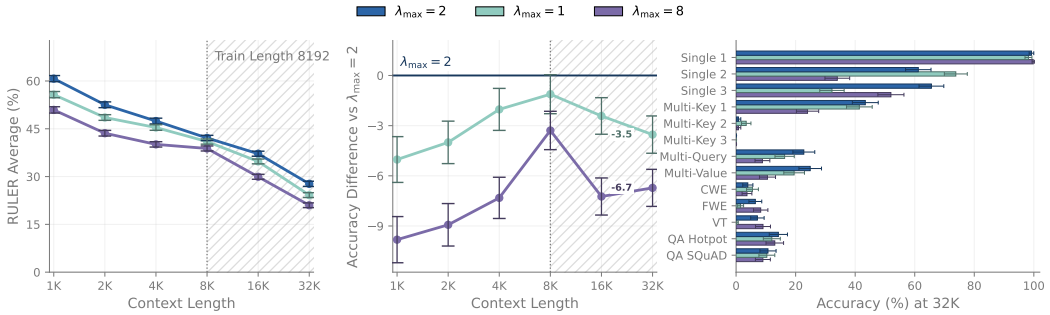

Figure 12: Ceiling of the trained speed, $\lambda_{\max} = 1$ and $\lambda_{\max} = 8$ against $\lambda_{\max} = 2$, all with $a_h \sim \mathcal{N}(0,1)$. Panels as in Fig. 11.

In this ablation, we train three 0.5B models with the same trainable resolution parameter and random initialization, but we vary $\lambda_{\max}$ over 1, 2 and 8. When $\lambda_{\max} = 1$, the most the model can rotate the fastest frequency is the base log-ratio model. When $\lambda_{\max} = 2$ or 8, the model can exceed the resolution of the standard log-ratio RoPE. Similar to Sec. F.4, we see that raising $\lambda_{\max}$ from 1 to 2 helps inside its training length, while limiting $\lambda_{\max} = 1$ limits both resolution and does not outperform $\lambda_{\max} = 2$ beyond the training length. As expected, allowing the resolution to go much higher hurts length extrapolation. $\lambda_{\max} = 8$ trails $\lambda_{\max} = 2$ at every length, and the gap grows from 3.3 points at 8K to 7.2 points at 16K.

G Extended Related Work

Positional encodings. The original transformer adds a sinusoidal or learned vector for each absolute position to the token embedding (Vaswani et al., 2023). Relative encodings instead give attention a function of the offset between the query and the key. Shaw et al. (2018) add a learned vector for each clipped offset, Transformer-XL adds a sinusoidal embedding of the offset to the attention score (Dai et al., 2019), and T5 learns a scalar bias for each offset bucket (Raffel et al., 2023). ALiBi (Press et al., 2022), KERPLE (Chi et al., 2022) and FIRE (Li et al., 2024) also add a bias to the logit. RoPE rotates the query and the key by their own positions (Su et al., 2023).

Proposed properties of an ideal positional encoding. Several works state properties that a positional encoding should have, and these properties often add a strong inductive bias. ALiBi adds a

penalty that grows linearly with distance, which biases attention towards recent tokens (Press et al., 2022). Su et al. (2023) motivate RoPE with a long-term decay, where the query–key inner product shrinks as the relative distance grows, but Barbero et al. (2025) show that this decay holds only in special cases and is unlikely to explain RoPE's success in a trained model. Sun et al. (2022) go further and argue that for a query q and two earlier positions $i < j$, $\mathbb{E}[\langle q, k_j \rangle] > \mathbb{E}[\langle q, k_i \rangle]$, so that position fixes a monotonic order of expected attention logits. The properties in Sec. 5 do not require a bias towards recent tokens or a fixed order of the logits.

Positional information without positional encoding. Haviv et al. (2022) show that causal language models without positional encoding reach perplexity close to models with one, and that a probe recovers position from their hidden states. Chi et al. (2023) trace this information to the variance of the attention output, which shrinks as a query attends to more keys. Kazemnejad et al. (2023) find that NoPE extrapolates better than several positional encodings on synthetic tasks.

RoPE beyond the training length. Men et al. (2024) and Liu et al. (2024) relate the usable context to the RoPE base and to the frequencies that do not complete a rotation within the training length. Position interpolation scales the positions down so that a target length maps into the trained range (Chen et al., 2023), and YaRN interpolates each frequency by a different amount (Peng et al., 2026). Lu et al. (2024) compare these methods under a controlled protocol, in which every method starts from the same base model and is trained on the same extension data. They find that exact-attention methods with fine-tuning are effective up to the length they were extended to. Beyond that length, extrapolation remains hard. Their study starts from a pretrained RoPE model and varies the extension method, whereas we vary the architecture during pretraining.

Hybrid attention. Gemma 2 and Gemma 3 interleave sliding-window and global layers to reduce the cost of long contexts (Team et al., 2024; 2025). Command A (Cohere et al., 2025), RNoPE (Yang et al., 2025a) and Muse-Glimmer-30B (Meta Superintelligence Lab, 2026) also change the positional encoding, with RoPE in the sliding-window layers and NoPE in the global layers. Yang et al. (2025a) find that NoPE layers attend sharply to retrieved tokens, while RoPE layers have a stronger preference for recent tokens.

H Limitations and Future Work

Our analysis uses 64M-parameter models, and the 0.5B models are trained with one learning rate at one training length. We have not tested whether the gap between Log-Ratio RoPE and RoPE persists at larger scales. Theorem 1 is a worst-case bound for keys with identical content. We study only softmax attention. Linear attention models such as Gated DeltaNet (Yang et al., 2025b) are also efficient on long sequences, like sliding-window attention, but we do not analyze them. We varied only the architecture. The training data, optimizer, and objective can also limit long-context performance, which we leave to future work.