ROPE DISTINGUISHES NEITHER POSITIONS NOR TOKENS IN LONG CONTEXTS, PROVABLY

Anonymous authors
Paper under double-blind review

ABSTRACT

We identify intrinsic limitations of Rotary Positional Embeddings (RoPE) in Transformer-based long-context language models. Our theoretical analysis is *only* related to the length of the context instead of its specific content. We prove that as context length increases, RoPE-based attention becomes unpredictable and loses two properties that are central to its effectiveness. First, it loses its locality bias: RoPE is no more likely to favor nearer positions than substantially farther ones. Second, it loses consistency in token relevance: a key vector that receives a higher attention score than an alternative at one position may receive a lower score at another. With a longer context, both failure modes approach a probability of 0.5, no better than random guessing. We further prove that the attention score can remain unchanged when a key token is moved to a different position, or even replaced by a different token, indicating a failure to distinguish positions or tokens. Increasing the RoPE base hyperparameter, a common practice in today's long-context models, helps distinguish different tokens, but inevitably sacrifices the ability to distinguish positions, as modifying the RoPE base merely trades off the two objectives but cannot preserve both at the same time. Our empirical analysis shows that multi-head, multi-layer architectures are insufficient to overcome these limitations. Our findings suggest that fundamentally new mechanisms for encoding position and token order may be needed in future Transformer long-context language models.

1 INTRODUCTION

Positional embeddings are essential in Transformers, because attention is otherwise permutation-invariant and cannot distinguish token order (Vaswani et al., 2017). Among the many positional embeddings, Rotary Positional Embedding (RoPE, Su et al., 2021) has emerged as the *de facto* choice in modern Transformer-based large language models (LLMs). This may arise from several attractive properties of RoPE. Through rotary operations, RoPE encodes the relative distance between tokens, and induces a locality bias that favors nearby tokens over distant ones (Su et al., 2021). Such inductive biases align well with the structure of natural language and prove beneficial for both training convergence (Gelberg et al., 2025) and extension to longer context lengths (Press et al., 2022).

Despite the increasing advertised context lengths of recent LLMs (Fu et al., 2024a; Team et al., 2024; DeepSeek-AI, 2026), many recent studies show that these models often struggle with long-context tasks they are supposed to be capable of, even at input lengths well within their claimed context lengths (Liu et al., 2024; Hsieh et al., 2024; Kuratov et al., 2024; Du et al., 2025). These recurring failures beg a fundamental question: *Are these failures artifacts of engineering choices, or do they reflect intrinsic limitations of RoPE itself?* This question is crucial in that it determines whether future progress in long-context Transformers should focus primarily on improved engineering, or instead require fundamentally new mechanisms for encoding positions and token order.

Our answer is that RoPE itself has intrinsic limitations in long contexts. We systematically explain this with a theoretical analysis of *single-head attention* that abstracts away from the specific content of the context and depends *only* on its length.[1] We show under mild assumptions[2] that as context

[1] We provide a discussion for the multihead and multilayer case in Section E.

[2] See §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.

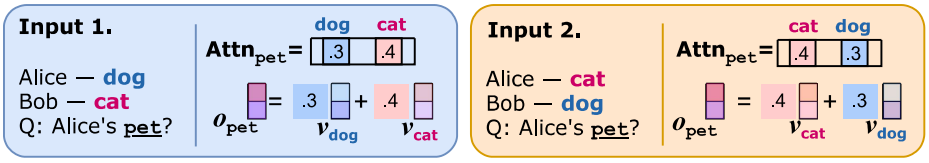


Figure 1: Position aliasing induces an attention invariance failure: there exist large numbers of positions where swapping two key tokens (`dog`, `cat`) keeps the attention output of a query token $o_{\texttt{pet}}$ unchanged.

length increases, RoPE's effect on attention becomes increasingly unpredictable and undermines the very properties that make it effective in language models, struggling on two primary objectives:

- First, RoPE fails to distinguish positions (§3). As the context length grows, the same token may receive a higher attention score at a farther position than at a closer one, with probability approaching 0.5 (**position inversion**; §3.1). RoPE thus becomes no better than random chance at favoring nearer positions over farther ones, effectively losing its locality inductive bias. We further identify a specific failure mode, which we call **position aliasing**: for a fixed query and key, moving the key to a different position may leave its attention score unchanged, so the model no longer distinguishes positions reliably (§3.2; Fig. 1).
- Second, RoPE fails to distinguish tokens (§4). As the context length grows, the relative ranking of two different key tokens for a given query, reflected by the attention scores they receive, can be arbitrarily reversed across positions: a token ranked above another at one position may be ranked below it at another (**token inversion**; §4.1). The probability of token inversion also approaches 0.5, no better than random chance. Moreover, longer context induces a phenomenon we call **token aliasing**: for a fixed query and key position, replacing the key token with a different token may leave the attention score unchanged, so the model effectively fails to distinguish tokens reliably (§4.2).

The above theoretical results are derived from a key new insight in our analysis, which treats the unnormalized attention score as a normal random variable (§2.2).

Our empirical analysis on Llama 3.1-8B (Grattafiori et al., 2024), which has a claimed context length of 128K tokens, confirms our theoretical conclusions about position and token inversion. It further shows that both position aliasing and token aliasing occur ubiquitously: across a context length of only 8K tokens, a staggering 75K pairs of positions exhibit position aliasing, appearing regardless of positional proximity; additionally, around 150 positions exhibit token aliasing in this range. Our theory suggests that commonly used length-extension techniques do *not* resolve the problem. Adjusting the RoPE base hyperparameter trades off the two failure modes rather than eliminating them. In particular, increasing the RoPE base helps preserve consistency in token relevance, but weakens the ability to distinguish positions.

Our experiments confirm that these failures persist in real multihead, multilayer LLMs (§5). We tested 6 popular models from 7B to over 100B on a simple task: given a list, the model must identify the value at the k-th position. This task addresses the ability to distinguish position, rather than distinguish token identities, since modern LLMs are commonly optimized for the latter through retrieval-style objectives (Kamradt, 2023). With just 4 distinct values in the list, all models perform no better than random guesses in as short as 4K tokens, a length disproportional to what these models were trained on. This strengthens our theoretical analysis of the single-head case by showing that the same positional failure persists in practical models.

Our findings temper some of the recent optimism created by rapidly increasing advertised context lengths. Extending the nominal context length alone is flawed if the underlying positional mechanism degrades as the context length grows. Our analysis provides a mechanistic explanation for the recurring long-context failures observed in recent studies (Liu et al., 2024; Hsieh et al., 2024; Kuratov et al., 2024; Du et al., 2025), suggesting that the gap between the nominal context limit and reliable use of distant information may not be eliminated through better data or engineering alone; instead, they reflect the fundamental limitations of the positional mechanism. By identifying

such limitations, this work motivates further study into fundamentally new approaches to positional mechanisms better suited to long-context language modeling.

2 DEMYSTIFYING ROPE

Attention in transformers should achieve two objectives: (1) **Position identification**, to encode where a token occurs in the text and allow attention to distinguish positions and capture contextual dependencies shaped by word order. Failures hurt the model's ability to understand the context dependency in tasks like counting or reasoning. (2) **Token identification**, to have each query distinguish among tokens and identify those that are contextually salient. Failures cause the model to ignore relevant inputs and generate hallucinated content. Long-context tasks often require a combination of these two objectives (Vaswani et al., 2017; Liu et al., 2024; Bai et al., 2024).

We define the **RoPE product** as the un-normalized attention score, i.e. the dot product between a query and a key after RoPE has been applied to both. This section aims to address two questions through the lens of the RoPE product: How does the RoPE product help with position and token identification (§2.1)? How does RoPE-based attention behave as the context length increases (§2.2)? We answer both through our key insight of treating the RoPE product as a normal random variable.

Throughout the paper, our theoretical analysis abstracts away from the specific content of the context and considers its length alone.

2.1 BACKGROUND

For a pair of query and key vectors $\mathbf{q}$ and $\mathbf{k}$, RoPE (Su et al., 2021) divides the d hidden dimensions into $h = d/2$ pairs of 2D vectors. As the token position changes, each 2D vector rotates at an angular frequency that is distinct to its dimension pair. The dot product between $\mathbf{q}$ and $\mathbf{k}$ after applying RoPE to both (the RoPE product) can be written as a function of their relative distance, m:

$$S(m) = S_{\mathbf{q},\mathbf{k}}(m) = \sum_{n=0}^{h-1} a_n \cos(m\theta^n + \phi_n).$$

The base frequency is $\theta = B^{-1/h} \in (0, 1)$ where $\Theta(B) > M$ is the RoPE base.[3] Vectors $\mathbf{a}$ and $\boldsymbol{\phi}$ are determined solely by $\mathbf{q}$ and $\mathbf{k}$. For the n-th frequency component, its amplitude $a_n > 0$ is the product of the norms of the corresponding 2D vectors (q_{2n}, q_{2n+1}) and (k_{2n}, k_{2n+1}), and its phase $\phi_n \in [0, 2\pi)$ is the angle subtended by them.

High-frequency components oscillate; low-frequency components decay For a context length limit of M, one typical way of analyzing the RoPE product is to separate the high and low frequency components using the threshold value $\lambda(M) = \Theta(h \log_B M)$ (Jonasson, 2025; Liu et al., 2023b; Peng et al., 2024; Miranda et al., 2024). For $m \in [0, M)$, **high-frequency** components complete at least one circle around the origin with $n \ll \lambda(M)$; **low-frequency** ones only rotate a small angle with $n \gg \lambda(M)$.

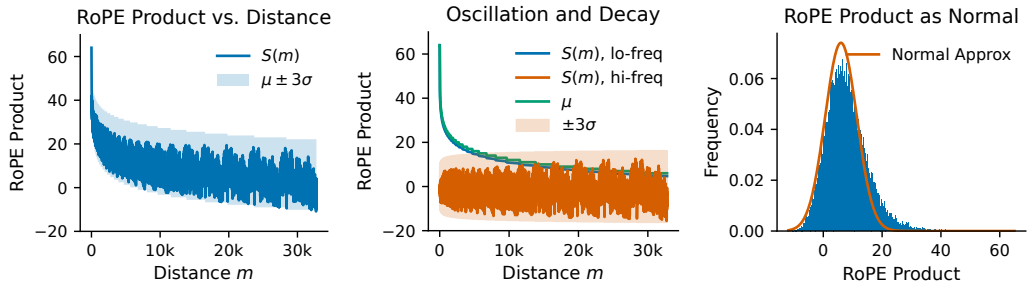
Fig. 2b illustrates the oscillation of the high-frequency components and the decay effect of the low-frequency ones.[4] RoPE helps with the two primary objectives discussed earlier. For position identification, high-frequency oscillation helps capture the difference between close positions, while the low-frequency decay globally distinguishes distant position pairs, promoting a locality inductive bias. For token identification, low-frequency components play a stabilizing role: their slower rotations preserve the relative ordering of token relevance, as they are less perturbed by relative distances.

2.2 KEY INSIGHT: THE ROPE PRODUCT AS A NORMAL RANDOM VARIABLE

Previous work has largely focused on low-frequency decay due to its analytical tractability (Miranda et al., 2024; Xu et al., 2024; Xiong et al., 2024). We develop a probabilistic characterization of

[3] Following standard practice we assume $M < \Theta(B)$, since otherwise even the lowest frequency term oscillates and loses its uniqueness (Liu, 2026) The factor can be 2π, π or other values under different criteria.

[4] Decay is not inevitable but is iconic in well-trained models. See Section A for a more in-depth discussion.



(a) Overall pattern of RoPE waveform. Shadow shows estimation of $S(m)$ as a normal.

(b) Low and high frequency parts of $S(m)$, and mean and standard deviation as a normal.

(c) Full distribution of $S(m)$ over $m \in [0, M)$.

Figure 2: Illustration of the RoPE product and its normal approximation when $S(m) = \sum_n \cos(m\theta^n)$, $m \in [0, 32, 768)$, $h = 64$, $B = 10^5$. 1k = 1,000.

the distributional behavior of the RoPE product. This perspective yields a deeper understanding of RoPE's behavior. A core theoretical contribution of this paper can be informally stated as follows:

Remark 2.1. If the distance m between a query $\mathbf{q}$ and a key $\mathbf{k}$ is randomly sampled from any interval $[A, M)$, where $M - A$ is large, then the RoPE product $S_{\mathbf{q},\mathbf{k}}(m)$ can be modeled as a normal variable

$$\widetilde{S} = \widetilde{S}_{[A,M)}(\mathbf{q}, \mathbf{k}) \sim N(\mu_M(\mathbf{q}, \mathbf{k}), \sigma_M^2(\mathbf{q}, \mathbf{k})),$$

with its mean decided by its low frequency terms, and its variance decided by its high frequency terms. The high frequency threshold is determined by the context limit, M.

Remark 2.1 follows from an application of the Central Limit Theorem. See §B for details and Fig. 2c for empirical validation. Remark 2.1 provides a powerful tool to characterize the behavior of RoPE product $S(m)$: it behaves approximately as a normal variable whose mean decreases (decay) and variance increases (oscillation) as the context length grows.

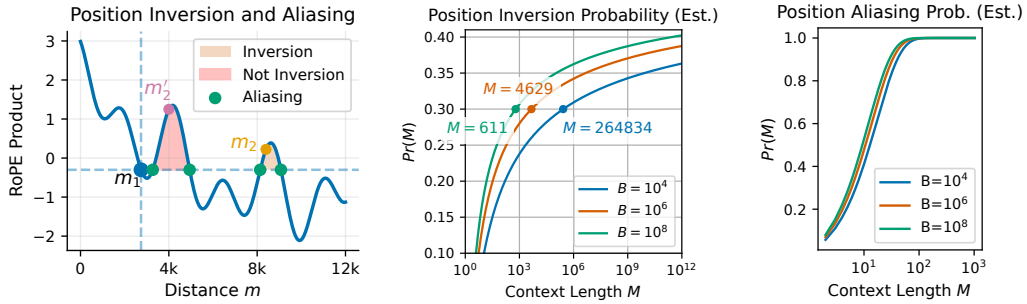
Organization of the rest of the paper The rest of the paper formalizes how RoPE's intrinsic properties undermine the fundamental objectives of both *position* and *token* identification in long contexts. We begin with a theoretical analysis of a single attention head under mild assumptions (see §C for details) in §3 and §4, where four specific failure modes are identified. For each, we first present a theoretical result and then provide empirical verification. Our empirical analysis probes an attention head from Llama 3.1-8B (Grattafiori et al., 2024), with a 128K claimed context length. We choose this model because of its popularity, moderate size, and representative decoder-only architecture.[5] We illustrate the failure modes with a long context of mostly irrelevant text containing three relevant sentences: "Alice has a cat," "Bob has a dog," and "What pet does Alice keep?" We analyze the key tokens "cat" and "dog" and the query token "pet". We use the first head in the first layer as a case study, although our method applies to any head in any layer. See Section D.1 for implementation details. In §5, we then turn to an empirical study of full multi-head, multi-layer language models.

3 RoPE Fails to Distinguish Positions in Long Contexts

For the position identification objective, suppose that we are given a pair of fixed query and key tokens in an input of length M. The tokens may be placed at any position as long as the query token appears later. This means that the relative distance m between the token pair satisfies $0 \leq m < M$.

With recency bias, we expect that the key should have a high chance of receiving larger attention weights when it is closer than when the same key token is located farther away (i.e. $S(m_1) > S(m_2)$ where $m_1 < m_2$). We identify two failure modes that violate this expected behavior and explain why they can be problematic.

[5] Although Llama 3.1 uses RoPE scaling, we show in Section B.1 that the analysis for standard RoPE still applies.



(a) Position inversion occurs if $S(m_2) > S(m_1)$ despite $m_2 \gg m_1$. Local oscillations (m_1, m_2') are *not* considered.

(b) Probability estimation (lower-bound) of position inversion vs. context length.

(c) Probability of position aliasing at a random distance vs. context length.

Figure 3: Illustrations (a) for position inversion and aliasing, with corresponding probability estimations under different RoPE **B**ases (b, c), $h = 64$. 1k = 1,000.

3.1 Failure Mode 1: Position Inversion

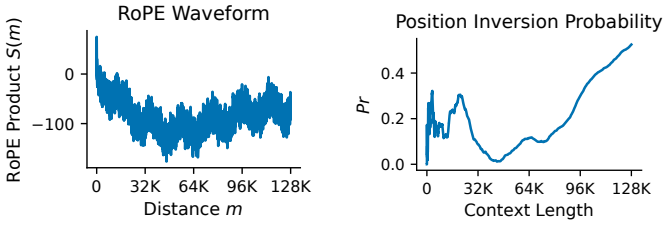
Position inversion is a reversal of RoPE's locality inductive bias: given the query, moving the key to a *substantially farther* position *increases* the attention score ($S(m_1) < S(m_2), m_1 < M/2 \leq m_2$). We draw inversions from distant pairs located on opposite halves of the context, since they are more detrimental than those among nearby tokens. See Fig. 3a for an illustrative example.

Theorem 1. *The probability lowerbound of position inversion increases with context length M and RoPE base B. The probability approaches $1/2$ as $\log M \log B \to \infty$.*

Theorem 1 follows directly from treating the RoPE product as a normal random variable, as discussed in Remark 2.1. See §C.1 for the formal statement and proof.

Theorem 1 states that, given a query, moving the exact key token from a closer position $m_1 \in [0, M/2)$ to a substantially farther position $m_2 \in [M/2, M)$ can increase its attention score with probability approaching that of a coin flip. This is problematic because, as the context length and RoPE base grow, attention becomes nearly arbitrary in its preference between nearby vs. farther positions, making its behaviors less predictable. This unpredictability may prevent the model from identifying a reliable positional pattern.

In practice, the probability of position inversion can exceed 0.3 even at short context lengths, as shown in Fig. 3b. Following a convention used in the Turing Test (Turing, 1950), we assume that this rate is already high enough to signal substantial positional ambiguity.



(a) The RoPE decay happens only within the initial $\sim$ 50K tokens.

(b) Position inversion probability approaches 0.5 as length increases.

Figure 4: Position inversion for key "cat" and query "pet". Llama 3.1-8B, Layer 0, Head 0. Here 1K = 1,024.

Empirical verification As shown in Fig. 4a, for the query token "`pet`", moving the key token "`cat`" across the advertised 128K context length of Llama 3.1-8B causes the attention score, i.e., the RoPE product, to reach a minimum at $m \approx 50$K. Beyond this point, the RoPE product exhibits an overall upward trend with oscillations, indicating position inversion.

Fig. 4b shows the corresponding probability of position inversion. Within just a few thousand tokens, this probability increases to nearly 0.3; once $m > 50$K, it continues to increase towards 0.5. Note again that we consider only pairs (m_1, m_2) where m_1 and m_2 lie in opposite halves of the full

context. These inversions indicate that the model can fail to properly compare a nearby token with a substantially farther one.

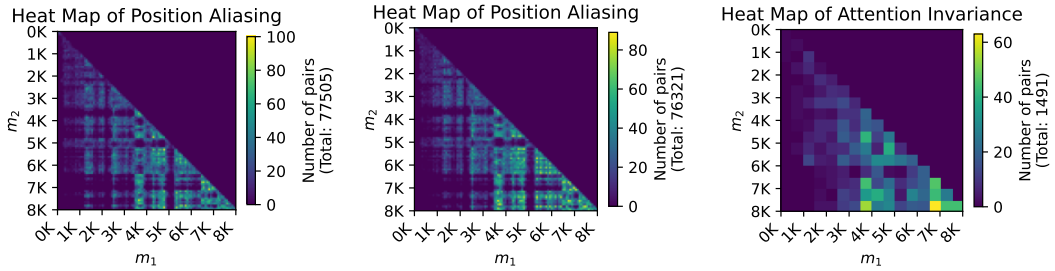
3.2 Failure Mode 2: Position Aliasing

Position aliasing occurs when modifying the distance between query and key does not change the attention score at all. Position aliasing can be seen as a complete failure to distinguish two different positions. Fig. 3a provides an illustration. An *aliasing pair* refers to two distances with the same attention score.

Theorem 2. *The probability that a random distance admits an aliasing pair converges to* 1 *exponentially fast as the context length* M *increases. Moreover, the total number of aliasing pairs increases with both the context length* M *and the RoPE base* B*.*

The intuition behind Theorem 2 is that the difference between the RoPE products at two independent positions can be modeled as a zero-mean normal random variable. This allows us to estimate how often its absolute value falls below the datatype resolution used for the RoPE product. See §C.2 for the formal statement and proof. See Fig. 3c for the probability estimation.

Theorem 2 states that position aliasing is inevitable with increased context lengths. In practice, the issue can be amplified by limited numerical precision: it occurs when the difference between two RoPE products falls below the resolution limit of the data type. Even when the attention scores for two distances are not exactly identical under higher precision, very small differences may be lost due to limited numerical precision.



(a) Position aliasing pairs for key "cat" and query "pet".

(b) Position aliasing pairs for key "dog" and query "pet".

(c) Attention invariance pairs for "cat", "dog" and "pet".

Figure 5: Heat maps of position aliasing and attention invariance pairs under BF16, showing the ubiquity of position aliasing. Llama 3.1-8B, Layer 0 Head 0. Pairs are grouped into a total of 200×200 bins for position aliasing, and 16×16 bins for attention invariance. 1K = 1,024.

Empirical verification As shown in Figs. 5a and 5b, under an 8K context length and commonly-used BF16 precision, almost every distance m_1 is involved in at least one aliasing pair (m_1, m_2), and there are already more than 75k aliasing pairs, the density of which increasing with the context length. This empirically confirms Theorem 2 and suggests that position aliasing is a common issue even at relatively short context lengths.

Attention invariance caused by position aliasing Position aliasing implies a specific failure mode: given a query $\mathbf{q}$ and two keys $\mathbf{k}_1$ and $\mathbf{k}_2$ at aliasing positions, swapping $\mathbf{k}_1$ and $\mathbf{k}_2$ does not change the attention output at all, as illustrated in Fig. 1. Fig. 5c empirically verifies this failure mode, showing 1,491 such invariance cases even within an 8K context length. This further demonstrates that position aliasing can be damaging even at short context lengths.

4 RoPE Gets in the Way of Distinguishing Tokens in Long Contexts

A good positional encoding is designed for position identification, but should not negatively affect token identification. Let $\mathbf{q}$ be a query vector and let $\mathbf{k}_1$ and $\mathbf{k}_2$ be two key vectors. We consider

the relative distances between the query and keys alone, but not a specific input context. Let $S_1(m)$ denote the RoPE product between $\mathbf{q}$ and $\mathbf{k}_1$ at distance m, and let $S_2(m)$ denote the corresponding RoPE product between $\mathbf{q}$ and $\mathbf{k}_2$. Assume that at $m = 0$, where RoPE effectively has no effect, the first key is more relevant, i.e. $S_1(0) > S_2(0)$. Intuitively, this relevance ordering is supposed to be preserved when both keys are placed at a new relative distance m, i.e. $S_1(m) > S_2(m)$. We identify the following violations.

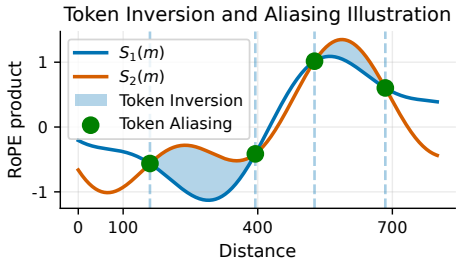
4.1 Failure Mode 3: Token Inversion

Token inversion occurs when the relevance ordering of the two keys is reversed at distance m, i.e. $S_1(m) - S_2(m) < 0$ despite $S_1(0) > S_2(0)$ (See Fig. 6a).

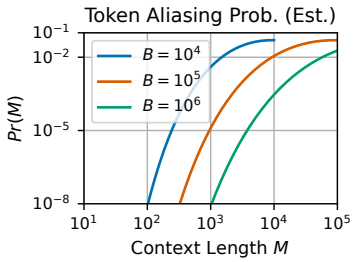
Theorem 3. *The probability lower bound for token inversion increases with the context length M, approaching $1/2$ as M approaches the natural context limit $\Theta(B)$. In contrast, the lower bound decreases with the RoPE base B.*

See §C.3 for the formal statement and proof.

Theorem 3 states that RoPE can reverse the original ordering between two keys at some nonzero relative distance. Similarly to position inversion (§3.1), the main problem with token inversion is its unpredictability: it can occur with probability approaching that of a coin flip. Suppose that $S_1(m) > S_2(m)$ for some values of m but $S_1(m) < S_2(m)$ for others, with comparable frequencies; then it becomes unclear whether the model can reliably distinguish the two keys at those distances.

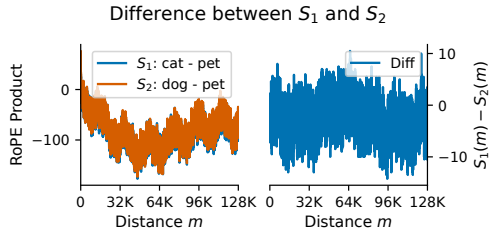


(a) Token inversion: $S_1(0) > S_2(0)$ but $S_1(m) < S_2(m)$. Token aliasing: $S_1(m) = S_2(m)$.

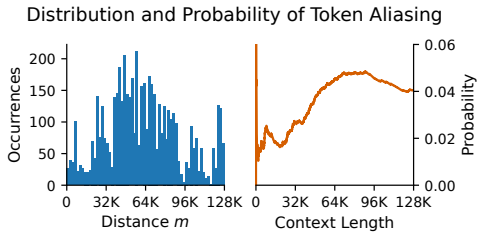


(b) Typical token aliasing probabilities at different RoPE **B**ases, using BF16.

Figure 6: Illustration of token inversion and aliasing.



(a) RoPE products for two key - query pairs (left) and their difference (right).



(b) Distribution (left) and probability (right) of token aliasing.

Figure 7: Token aliasing probabilities under BF16, for query `pet`, and keys `cat` (S_1) and `dog` (S_2). Llama 3.1-8B, Layer 0 Head 0. Here, 1K = 1,024.

Empirical verification For the query token `pet`, we select a highly relevant key token, `cat`, and a less relevant key token, `number`. Let S_1 denote the RoPE product between `pet` and `cat`, and let S_2 denote the RoPE product between `pet` and `number`. Fig. 8 shows the difference $D = S_1 - S_2$ and the probability curve of token inversion. Initially, $D > 0$, as desired, indicating that `cat` receives a higher score than `number`. However, in fewer than 10 tokens, D drops below zero and the relevance ordering between the two tokens is already reversed. As m increases, the probability of inversion

exhibits an increasing lower bound, consistent with Theorem 3. When $m \geq 20K$, the probability approaches 0.5; with an oscillating D, it becomes unpredictable whether `cat` or `number` receives the higher attention score.

4.2 Failure Mode 4: Token Aliasing

At relative distance m, replacing $\mathbf{k}_1$ with a different key $\mathbf{k}_2$ can leave the attention score unchanged, i.e., $S_1(m) = S_2(m)$. We refer to this phenomenon as **token aliasing**, which indicates that attention fails to distinguish between two different tokens at that position, as illustrated in Fig. 6a.

Theorem 4. *The number of token aliasing positions increases with M and decreases with B. For a sufficiently long context of length M, it is bounded by $\Theta(2^{-f}\sqrt{h}M)$, where f is the explicit fraction bits of the data type used, and h is the half hidden dimension.*

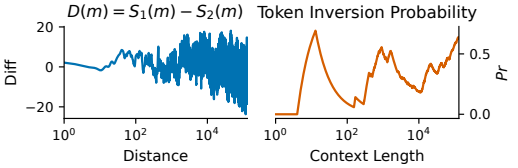


Figure 8: Token inversion for keys "cat", "number" and query "pet". Left: difference of RoPE product vs. distance, where S_1 corresponds to "cat", and S_2 to "number". Right: Probability. Llama 3.1-8B, Layer 0 Head 0.

See §C.4 for the formal statement and proof. Similar to position aliasing, token aliasing is amplified by limited numerical precision, such as BF16 where there are $f = 7$ explicit fraction bits (Henry et al., 2019).

Table 1: Summary of failure modes and how the chances of occurrence change with context length M and RoPE base B. It is preferable to have smaller probabilities as opposed to larger ones: for inversion this means better predictability, and for aliasing this means less ambiguity.

Failure Mode	Indicator	$M \uparrow$	$B \uparrow$
Position Inversion	$m_1 < m_2; S(m_1) < S(m_2)$	↑	↑
Position Aliasing	$S(m_1) = S(m_2)$	↑	↑
Token Inversion	$S_1(0) > S_2(0); S_1(m) < S_2(m)$	↑	↓
Token Aliasing	$S_1(m) = S_2(m)$	↑	↓

Token aliasing can be mitigated by increasing RoPE base (see Fig. 6b); nevertheless, it is almost always present in long inputs. If $h = 64$, using BF16, up to 5% of the total positions exhibit token aliasing. This means 1.6K aliasing positions in 32K tokens.

Empirical verification For query `pet` and two keys `cat` and `dog`, Fig. 7a shows the difference between the two corresponding RoPE products $D(m) = S_1(m) - S_2(m)$. Under BF16 precision, we count the aliasing positions where $D(m) = 0$ and calculate the aliasing probability (Fig. 7b). According to Theorem 4, the probability should converge to a value of around 0.05, which matches the illustrated result. Fig. 7b left shows that the frequency of token aliasing is negatively correlated with the RoPE products, which demonstrates how decay affects token aliasing.

We close §3 and §4 with Table 1 summarizing the four failure modes, along with the following takeaways:

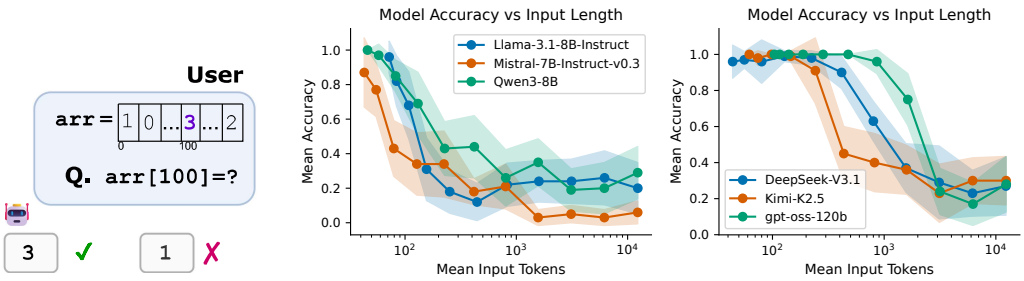
Takeaways

As the context length increases, all four failure modes become more likely, and RoPE-based attention becomes increasingly likely to fail at distinguishing both positions and tokens. The choice of RoPE base B trades off the failure modes: While increasing RoPE base can mitigate token inversion and token aliasing, it does not fully resolve them; moreover, it worsens position inversion and position aliasing. In this sense, each attention head has a maximum effective context length. Beyond this limit, at least some of the four failure modes occur with sufficiently high frequency and compromise attention, regardless of how RoPE base is adjusted.

5 How Do Multilayer, Multihead Transformer LLMs Fare?

We have shown the four failure modes of a single attention head. But do multilayer, multihead Transformer LLMs overcome the limitations in practice? This section addresses this empirically.

We conduct a controlled evaluation of six open RoPE-based long-context LLMs of different sizes, as shown in Fig. 9. We do not evaluate closed-source proprietary models, since their architectures and positional-embedding choices are not publicly known. Moreover, many recent long-context models are explicitly optimized for retrieval-based tasks, such as Needle-in-a-Haystack (Kamradt, 2023), which are essentially token-identification tasks. Our theoretical results predict that optimizing for distinguishing tokens inevitably trades off against distinguishing positions; therefore, we focus on the latter through controlled experiments.



(a) The indexing task. (b) Small models ($\leq$ 8B). (c) Large models ($\geq$ 100B).

Figure 9: (a) We ask models to extract the k-th element in an array consisting of only integers 0-3. (b, c) Dots and shadows represent mean and standard deviation of accuracy. Selected models: Grattafiori et al. (2024); Mistral AI Team (2024); Yang et al. (2025a); DeepSeek-AI et al. (2025); Team et al. (2026a); OpenAI et al. (2025)

The indexing task In our position identification task, the model is presented with a Python list `arr`, and is required to answer the value of the given index `arr[i]`. Each model receives the same input. For each list length, we test the models with 100 samples. The list only contains integers 0, 1, 2, and 3. This turns the task into a multiple-choice problem, allowing us to compare with the random-guess accuracy: if a model cannot identify the single target position (each element takes 1 token), it answers a random element with an accuracy of 0.25.

Results As shown in Fig. 9, all models start with near-perfect accuracy but quickly drop to a level close to random guessing. These real long-context models face serious position confusion with as short as several thousands of tokens, a cost they pay for optimizing token identification.

See §D for more details on the experiment settings.

6 Conclusion and Discussion

Our theoretic analysis shows that RoPE intrinsically fails to distinguish position identity in long inputs, which is its original purpose, and while it is not responsible for identifying tokens, RoPE nevertheless makes it worse as contexts get longer. The selection of RoPE base only acts as a tradeoff between the two objectives. The same bottleneck persists in practical multi-head multi-layer models, as observed in the empirical verifications.

Our conclusion suggests that a robust positional mechanism should maintain the ability to distinguish positions and tokens, and context length cannot be effectively extended without addressing the two objectives. At the same time, we are encouraged by recent efforts that explore alternatives to direct length extension, including improved context management and alternative paradigms such as recursive or agentic language models. Although these approaches do not directly resolve the intrinsic limitations of RoPE identified in this work, they reflect a broader shift away from treating context extension alone as sufficient. We hope that this work encourages further research into fundamentally new positional mechanisms for long-context language models.

AI USE STATEMENT

In this work, we used generative AI to help find related works on recent theoretic progress on RoPE and relevant mathematic tools. We also used generative AI and grammar checkers to proofread and polish the writing. We did *not* use AI to generate research ideas, write code, organize experiment data, draw figures, or draft paragraphs. We reviewed all related works provided by AI and validated the accuracy of the information. We made the final decision on the polished wording. We take full responsibility for the content of this work in its entirety.

REFERENCES

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.

Yuhan Chen, Ang Lv, Ting-En Lin, Changyu Chen, Yuchuan Wu, Fei Huang, Yongbin Li, and Rui Yan. Fortify the shortest stave in attention: Enhancing context awareness of large language models for effective tool-use. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11160–11174. Association for Computational Linguistics, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. Technical report, DeepSeek-AI, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf. Technical report.

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.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. Technical report, DeepSeek-AI, April 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf. Technical report.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context length alone hurts llm performance despite perfect retrieval, 2025. URL https://arxiv.org/abs/2510.05381.

Carl-Gustav Esseen. On the liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28: 1–19, 1942.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. *arXiv preprint arXiv:2402.10171*, 2024b.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7376–7399, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.366. URL https://aclanthology.org/2025.acl-long.366/.

Yoav Gelberg, Koshi Eguchi, Takuya Akiba, and Edoardo Cetin. Extending the context of pretrained llms by dropping their positional embeddings, 2025. URL https://arxiv.org/abs/2512.12167.

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.

Greg Henry, Ping Tak Peter Tang, and Alexander Heinecke. Leveraging the bfloat16 artificial intelligence datatype for higher-precision computations, 2019. URL https://arxiv.org/abs/1904.06376.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024. This work is licensed under the Apache License 2.0.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023.

André Jonasson. Rotary offset features in large language models, 2025. URL https://arxiv.org/abs/2503.01832.

Patrick Kahardipraja, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. The atlas of in-context learning: How attention heads shape in-context retrieval augmentation, 2025. URL https://arxiv.org/abs/2505.15807.

Gregory Kamradt. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/7b90d285651b68d39a94f3d3bd3672f84192c989, 2023.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack, 2024.

Shenggui Li, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li, and Yang You. Sequence parallelism: Long sequence training from system perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2391–2404, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.134. URL https://aclanthology.org/2023.acl-long.134/.

Yuping Lin, Zitao Li, Yue Xing, Pengfei He, Yingqian Cui, Yaliang Li, Bolin Ding, Jingren Zhou, and Jiliang Tang. Retrieval heads are dynamic, 2026. URL https://arxiv.org/abs/2602.11162.

Feilong Liu. Rotary positional embeddings as phase modulation: Theoretical bounds on the rope base for long-context transformers, 2026. URL https://arxiv.org/abs/2602.10959.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023a. URL https://arxiv.org/abs/2310.01889.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation. *arXiv preprint arXiv:2310.05209*, 2023b.

Magic. 100m token context windows, August 2024. URL https://magic.dev/blog/100m-token-context-windows.

Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

David Miranda et al. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024.

Mistral AI Team. Mistral-7b-instruct-v0.3. `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`, 2024. Model card. Accessed: 2026-05-03.

OpenAI. Introducing GPT-5, 2025. https://openai.com/index/introducing-gpt-5/.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL `https://arxiv.org/abs/2508.10925`.

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

Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=R8sQPpGCv0`.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia

Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael

Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam

Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong,

Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026a. URL https://arxiv.org/abs/2602.02276.

Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, Yutian Chen, Junjie Yan, Ming Wei, Y. Zhang, Fanqing Meng, Chao Hong, Xiaotong Xie, Shaowei Liu, Enzhe Lu, Yunpeng Tai, Yanru Chen, Xin Men, Haiqing Guo, Y. Charles, Haoyu Lu, Lin Sui, Jinguo Zhu, Zaida Zhou, Weiran He, Weixiao Huang, Xinran Xu, Yuzhi Wang, Guokun Lai, Yulun Du, Yuxin Wu, Zhilin Yang, and Xinyu Zhou. Attention residuals, 2026b.

Together AI. Together AI Docs. https://docs.together.ai/, n.d. Accessed: 2026-05-04.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. ISSN 00264423. URL http://www.jstor.org/stable/2251299.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling, 2023.

/u/bloc97. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation. https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/, June 2023. Reddit post on r/LocalLLaMA.

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

Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, Qi Zhang, Xuanjing Huang, and Xiaoling Wang. Length generalization of causal transformers without position encoding, 2024. URL https://arxiv.org/abs/2404.12224.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-demos.6.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality, 2024. URL https://arxiv.org/abs/2404.15574.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv*, 2023.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.260. URL https://aclanthology.org/2024.naacl-long.260/.

Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Base of rope bounds context length. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 87386–87410. Curran Associates, Inc., 2024. doi: 10.52202/079017-2773. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9f12dd32d552f3ad9eaa0e9dfec291be-Paper-Conference.pdf.

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

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, 2025b. URL https://arxiv.org/abs/2501.18795.

Meizhi Zhong, Chen Zhang, Yikun Lei, Xikai Liu, Yan Gao, Yao Hu, Kehai Chen, and Min Zhang. Understanding the rope extensions of long-context llms: An attention perspective, 2024. URL https://arxiv.org/abs/2406.13282.

A Rotary Positional Embedding

Let $X = (x_0, x_1, \ldots, x_{M-1})$ be the input word embeddings, where M is the number of input tokens, and $x_i \in \mathbb{R}^d$ be the word embedding of token i. The word embeddings are then transformed into

query, key and value representations, the transformation typically being linear, where the positional information is embedded.

RoPE applies the same multiplication-based positional embedding to the query vector $\mathbf{q}$ and key vector $\mathbf{k}$:

$$\hat{q}_i = R^d_{\Theta,i} q_i = R^d_{\Theta,i} W_Q x_i,$$
$$\hat{k}_j = R^d_{\Theta,j} k_j = R^d_{\Theta,j} W_K x_j,$$

where

$$R^d_{\Theta,m} = \begin{pmatrix} \cos m\theta^0 & -\sin m\theta^0 & 0 & 0 & \dots & 0 & 0 \\ \sin m\theta^0 & \cos m\theta^0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \cos m\theta^1 & -\sin m\theta^1 & \dots & 0 & 0 \\ 0 & 0 & \sin m\theta^1 & \cos m\theta^1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \cos m\theta^{d/2-1} & -\sin m\theta^{d/2-1} \\ 0 & 0 & 0 & 0 & \dots & \sin m\theta^{d/2-1} & \cos m\theta^{d/2-1} \end{pmatrix}.$$

The attention mechanism features the "attention score", defined as

$$A = \text{softmax}\left(\frac{\hat{Q}^\intercal \hat{K}}{\sqrt{d}}\right),$$

where $\hat{Q} = (\hat{q}_0, \hat{q}_1, \dots, \hat{q}_{M-1})$ and $\hat{K} = (\hat{k}_0, \hat{k}_1, \dots, \hat{k}_{M-1})$.

The attention score $a_{i,j}$ is the normalization of the inner product between $\hat{q}_i$ and $\hat{k}_j$. For the sake of simplicity, we call this inner product the "RoPE product". RoPE is designed such that only the relative distance, $i - j$, matters in RoPE product:

$$\hat{q}_i^\intercal \hat{k}_j = (R^d_{\Theta,i} q_i)^\intercal (R^d_{\Theta,j} k_j) = q_i R^d_{\Theta,i-j} k_j.$$

Conventions We use the following conventions throughout this paper:

- All indices start from 0;
- d — hidden dimension of an attention head;
- $h \equiv d/2$ — half dimension;
- B — the RoPE base (also called "the RoPE theta" in some literature);
- $\theta \equiv B^{-1/h}$ — the basic RoPE frequency (not "the RoPE theta");
- m — the distance between the query token, $\text{tok}_{\text{idx}_{\text{query}}}$, and the key token, $\text{tok}_{\text{idx}_{\text{key}}}$, defined as $m = \text{idx}_{\text{query}} - \text{idx}_{\text{key}}$. For causal models, $m \geq 0$;
- $\mathbf{q}, \mathbf{k}$ — the query and key vectors;
- $S_{\mathbf{q},\mathbf{k}}(m)$ — the RoPE product of the given query and key vectors, w.r.t. their distance m. By default, when referring to $\mathbf{q}$ and $\mathbf{k}$, we use the abbreviated $S(m)$;
- $\mathbf{a}, \boldsymbol{\phi}$ — the amplitudes and phases of the RoPE product, defined as

$$a_n = a_n(\mathbf{q}, \mathbf{k}) = \sqrt{(q_{2n}^2 + q_{2n+1}^2)(k_{2n}^2 + k_{2n+1}^2)},$$
$$\phi_n = \phi_n(\mathbf{q}, \mathbf{k}) = \text{atan2}(q_{2n}k_{2n+1} - q_{2n+1}k_{2n}, q_{2n}k_{2n} + q_{2n+1}k_{2n+1});$$

- $\lambda(M)$ — the threshold dimension $\lambda(M) = \Theta(h \log_B M)$.

Table 2: A summary of notations we use.

Notation	Description
d	Hidden dimension for an attention head
h	Number of rotary components $h = d/2$
B	RoPE base, the most important RoPE hyperparameter
θ	Base frequency $\theta = B^{-1/h}$
$\mathbf{q}, \mathbf{k} \in \mathbb{R}^d$	query and key vectors
M	Context length limit
$S(m)$	RoPE product of $\mathbf{q}$ and $\mathbf{k}$ at distance m
a_n, ϕ_n	Amplitude and phase of the n-th rotary component
$\lambda(M)$	Threshold frequency index

The High and Low Frequencies of RoPE Product We use the threshold value $\lambda = \lambda(M) = \Theta(h \log_B M)$ to separate high and low frequencies: the terms with smaller indices, $n \ll \lambda(M)$, are considered high frequency terms, and the others are low frequency terms. This is not a strict division; generally, low frequency terms should rotate no more than a complete circle within the context length limit.

The high frequency components correspond to the oscillation effect. Across the distance m, these rapid rotations create a distinction in the RoPE products of close position pairs.

The low frequency components cause the decaying effect (the recency bias feature). The low frequency terms rotate slowly, and are generally within the decaying interval. More specifically, when $n \gg \lambda(M)$, we have $m\theta^n + \phi_n \in [0, \pi]$, so $\cos(m\theta^n + \phi_n)$ decreases with m on $m \in [0, M)$. This decaying effect identifies distant positions. Fig. 2b illustrates both the oscillation and the decaying effect.

The low frequency components also help preserve token relevance. For a low frequency term $n \gg \lambda(M)$ with very little rotation, $m\theta^n + \phi_n \approx \phi_n$, and the cosine term changes little from its initial value. It has been shown (Miranda et al., 2024; Jonasson, 2025) that for query-key pairs that reflect longer text dependency, the low frequency terms tend to have higher amplitudes a_n, so that the RoPE product is less affected by the distance.

The Natural Context Length Limit If the context length M is too large, even the lowest frequency term begins to oscillate. From a signal processing perspective, the lowest frequency term $n = h - 1$ determines the fundamental frequency $\theta^{h-1} = B^{-\frac{h-1}{h}} \approx B^{-1}$, with a fundamental wavelength of $2\pi B$[6]. If $M \geq 2\pi B$, the fundamental frequency term loses its uniqueness and the positional embedding becomes ambiguous (Liu, 2026). Therefore, given a RoPE base B, $M < 2\pi B$ is a natural upperbound for context length M. If we expect RoPE to maintain recency bias, then the lowest frequencies must be in their decreasing phase, lowering this bound even further to around πB. In this paper, we use a rough estimation $\Theta(B)$ as the natural context length limit; by default, M does not exceed this limit.

Decay It is worth noticing that for the actual RoPE product $S(m)$, the decay is not universal. First, the decay is the collective effect of multiple low-frequency rotations, so it occurs only within an interval, after which even the lowest frequency starts oscillating. Second, it is possible that the RoPE product does not decay from the beginning; the initial decay only occurs when the low frequency phases ϕ_n are mostly small ($< \pi/2$). In fact, for any distance m, we may construct a phase vector $\boldsymbol{\phi}$ so that $S(m)$ reaches its maxima at m (Miranda et al., 2024) [7]. It is not uncommon in reality that $S(m)$ reaches its peak before starting the decay. However, as we show in Section 3, the decay is a preferred behavior: a lack of decay indicates that there is no global monotonicity, leading to a failure to identify distant positions (position inversion).

[6] Assuming that the lowest frequency amplitude $a_{h-1} > 0$.

[7] A more accurate statement on decay should be: if $\widetilde{S} \sim N(\mu_M, \sigma_M^2)$, then the upperbounds of the 99.7% probability threshold $\mu_M + 3\sigma_M$ decrease with $M \in [0, B^{h_1/h})$, where h_1 is the largest index such that $a_{h_1}/a_{\max} \gg 0$, i.e. the lowest frequency component with a non-negligible amplitude.

Non-Uniform Rotary Amplitudes We assume that the magnitudes $\{a_n\}$ across frequency terms are relatively uniform. When, as seen in real models, the low frequency terms instead have a negligible magnitude compared to high frequency ones, the actual base wavelength shrinks to match with the first term with a substantial magnitude. This in effect equals applying a smaller B: it shortens the natural context length limit, and causes a shorter decay interval.

B RoPE Product Can Be Seen as a Normal Variable

Consider the high frequency parts where $n < \lambda(m)$. For any integer $0 < A \leq h$, define the partial RoPE product as

$$S_{n<A}(m) = \sum_{n<A} a_n \cos(m\theta^n + \phi_n).$$

We shall show that if m is uniformly chosen on $[A, M)$ where $M - A$ is large, then $S_{n<\lambda(M)}(m)$ heuristically behaves like a normal variable. Then, by estimating the low dimension terms as $S_{n\geq\lambda(M)}(m) \approx \sum_{n\geq\lambda(M)} a_n \cos\phi_n$, we can estimate the distribution of the whole RoPE product.

Suppose that we are interested in the integer interval $I_A = [A, A + M)$, on which m is a uniform random variable. On I_A, define the mean exponential sum

$$\begin{aligned} G_M(\alpha, A) =& \frac{1}{M} \sum_{m=A}^{A+M-1} e^{i\alpha m} \\ =& \frac{\sin(M\alpha/2)}{M\sin(\alpha/2)} e^{i\alpha(A+(M-1)/2)}. \end{aligned}$$

We have

$$|G_M(\alpha, A)| \ll \min\left(1, \frac{1}{M\sin(\alpha/2)}\right).$$

For higher frequencies, $M \gg \frac{1}{\sin(\alpha/2)}$, and $|G_M(\alpha, A)| \ll 2/M\theta^n$.

The Moments: Expectation and Variance The mean cosine sum, $\mu_M(A)$, can be defined as

$$\begin{aligned} \mu_M(A) =& \frac{1}{M} \sum_{m=A}^{A+M-1} S(m) \\ =& \sum_{n=0}^{\lambda(m)-1} a_n \Re(e^{i\phi_n} G_M(\theta^n, A)). \end{aligned}$$

Let us first calculate the variances of individual frequency terms. For the weighted cosine value at the n-th dimension pair, the term $\Psi_n = a_n \cos(m\theta^n + \phi_n)$ is a random variable, with the following results:

$$\begin{aligned} E_m[\Psi_n] =& a_n \Re(e^{i\phi_n} G_M(\theta^n, A)) \approx 2a_n/M\theta^n \to 0, \\ E_m[\Psi_n^2] =& \frac{a_n^2}{2}\left(1 + \Re(e^{2i\phi_n} G_M(2\theta^n, A))\right) \approx \frac{a_n^2}{2}\left(1 + \frac{1}{4M\theta^n}\right) \to \frac{a_n^2}{2}. \end{aligned}$$

Next, let us estimate the cross variance terms. Let n, p be two frequency indices such that $n < p$. Using $\cos a \cos b = \frac{1}{2}\cos(a + b) + \frac{1}{2}\cos(a - b)$, and $\theta^n - \theta^p \asymp \theta^n + \theta^p \asymp \theta^n$,

$$|E_m[\Psi_n\Psi_p]| \ll \frac{a_n a_p}{M(1-\theta)\theta^n} \sim \frac{a_n a_p}{M\theta^n}(1/2 + \frac{h}{\log B}).$$

When $M\theta^n \gg 1$ ($h/\log B$ to be exact), the only significant terms related to the variance are the variants of individual cosine terms, and we may ignore the covariances, with $O(h/\log B)$ exceptions of size $O(a_n a_p h/\log B)$.

$$\begin{aligned} Var(S_{n<\lambda(M)}) &= \sum_{n<\lambda(M)} Var(\Psi_n) + \sum_{n<\lambda(M)} \sum_{n<p<\lambda(M)} Cov(\Psi_n, \Psi_p) \\ &\asymp \sum_n a_n^2/2 + o(\frac{\sum_{n,p} a_n a_p}{\log^2 B}). \end{aligned}$$

The second (covariance) term depends on $\frac{1}{1-\theta} = \Theta(h/\log B + 1/2)$. We shall proceed to show that this term may be neglected. In fact, if the numbers

$$\frac{1}{2\pi}, \frac{\theta}{2\pi}, \ldots, \frac{\theta^{\lambda(M)}}{2\pi} \tag{1}$$

are linearly independent over $\mathbb{Q}$ (i.e. for any $\mathbf{k} \in \mathbb{Q}^{\lambda(M)+1}$, $\sum_n k_n\theta^n/2\pi = 0$ iff $\mathbf{k} = \mathbf{0}$), then according to the multi-dimensional generalization of Weyl's Criterion, the vector sequence $\{\mathbf{v}_m\}$ is equidistributed modulo 1, where

$$\mathbf{v}_m = (\frac{m}{2\pi}, \frac{m\theta}{2\pi}, \ldots, \frac{m\theta^{\lambda(M)}}{2\pi}).$$

The independence of sequence 1 is equivalent to that of the sequence

$$1, \theta, \ldots, \theta^{\lambda(M)}.$$

Indeed, this sequence is linearly dependent over $\mathbb{Q}$ if and only if θ is a root of a rational polynomial of degree at most $\lambda(M) < h$.

Pseudo Independence Since $\theta = B^{-1/h}$, or equivalently $B\theta^h - 1 = 0$, where h is a positive integer, θ may be a root of such a polynomial only if B is a perfect power, i.e. $\exists g, c, \alpha, \beta, \gamma_i \in \mathbb{N}$ s.t. $B = c^\beta$, $c = \prod_i p_i^{\gamma_i}$, $g = \beta \cdot \gcd_i\{\gamma_i\}$ and $\gcd(g, h) > 1$. Actually, this can be easily constructed: an example is $B = 65536 = 2^{16}, h = 32$, where $\theta = 1/\sqrt{2}$ is a root with degree 2. However, even if θ^k is rational for some $k = h/\gcd(g, h)$, this only means that the sequence can be divided into $O(h/k) = O(\gcd(g, h))$ independent groups of size $O(k)$. One bound is that $k \geq 2$ whenever $B < 2^h$. For most cases where B is not a perfect power, $\gcd(g, h) = 1, k = h$. Therefore, we can heuristically view $X_j = m\theta^j$ as (largely) independent random variables that follow uniform distribution on $[0, 2\pi)$, with negligible covariances between any two terms.

What if k is small? We only need to study the covariances of the dependent pairs. Let n, p be the indices of a dependent pair where $n < p, n < \lambda(M), B = b^{h/k}, \theta = b^{-1/k}, b \geq 2, k = 2$ (the smallest k possible). Then $p = ckn$ for some integer $c \geq 1$.

$$\Psi_n\Psi_p = \frac{1}{2}a_n a_p(\cos(m\theta^n + m\theta^p + \phi_n + \phi_p) + \cos(m\theta^n - m\theta^p + \phi_n - \phi_p))$$

where $m\theta^n + m\theta^p = m\theta^n(1+\theta^{(ck-1)n}) \geq m\theta^n, m\theta^n - m\theta^p = m\theta^n(1-\theta^{(ck-1)n}) \geq m\theta^n(1-\theta^n)$, so

$$E_m[\Psi_n \Psi_p] \ll \frac{a_n a_p}{M\theta^n(1-\theta)} = \frac{a_n a_p}{M\theta^n(1-b^{-1/2})} \leq \frac{a_n a_p}{(1-1/\sqrt{2})M\theta^n} = o(a_n a_p).$$

If k is small, the covariances between dependent terms become much smaller than the variance terms ($\Theta(a_n^2)$) and may be ignored.

Normal Approximation Since $\{a_n\}$ is assumed with no dominating terms, we can apply Lindenberg's CLT and approximate the sum $S_{n<\lambda(M)}$ as a normal distributed random variable with zero mean.

The estimated distribution of the full RoPE sum is therefore:

$$\widetilde{S}_M \sim N(\mu_M, \sigma_M^2), \tag{2}$$

where

$$\mu \approx \sum_{n=\lambda(M)}^{h-1} a_n \cos\phi_n \leq \sum_{n=\lambda(M)}^{h-1} a_n, \tag{3}$$

$$\sigma \approx \sqrt{\sum_{n=0}^{\lambda(M)-1} \frac{a_n^2}{2}}. \tag{4}$$

Error Estimation Since CLT is used to derive the normal approximation, it is good when $\lambda(M)$ is large. Following the Berry-Esseen Theorem (Esseen, 1942), the distribution approximation error is on the order of $O\left(\frac{\rho}{\sigma^3\sqrt{\lambda(M)}}\right)$, where $\rho = E(|a_n \cos x_n|^3) = O(a_n^3)$. Therefore, the error is

$$O\left(\frac{1}{\sqrt{\lambda(M)}}\right).$$

Empirically, the practical convergence is usually better, and $\lambda(M) > 20$ is usually good enough. For $h = 64$, this means that the estimation is good when $M > \sqrt[3]{B}$.

B.1 RoPE Scaling

For any variant of RoPE, we need to analyze the linear independence of its frequencies (angular frequencies ω_n over 2π), i.e.

$$\frac{\omega_0}{2\pi}, \frac{\omega_1}{2\pi}, \ldots, \frac{\omega_{\lambda(M)}}{2\pi}$$

We study RoPE scalings where the n-th angular frequency is a polynomial of θ^n, i.e.

$$\omega_n = \sum_{p=0}^{E_n} \kappa_{n,p}\theta^{np}, \kappa_{n,E_n} > 0, E_n > 0.$$

This includes most RoPE scaling variants, such as the NTK scaling (/u/bloc97, 2023) used in Llama models.

If all $\kappa_{n,p}$ are rational, then such independence can be largely analyzed the same way: it can be secured if θ is not a root of any rational polynomial of $\lambda(M) \cdot \max\{E_n\}$. Again, this is rarely the case. If for some specific selections of B andh h, $\theta^{k \max E_n}$ is rational for some $k > 1$, then $1 - \theta$ should be $O(1)$, and the resulting covariances of the dependent terms should also be negligible.

The subsequent analyses are similar. The only major difference is that $\lambda(M)$ should be modified accordingly to satisfy $M\omega_{\lambda(M)} = \Theta(1)$.

C THE FAILURE MODES

C.1 POSITION INVERSION

Definition C.1. For a pair of query and key vectors $\mathbf{q}, \mathbf{k}$, context length limit M, uniformly select $m_1 \in [0, M/2)$ and $m_2 \in [M/2, M)$. A **position inversion** occurs for the pair (m_1, m_2) if and only if $S(m_1) < S(m_2)$. The probability of "position inversion" can be defined as

$$Pr(\text{inversion}|\theta, M, \mathbf{q}, \mathbf{k}) = Pr(S_{\mathbf{q},\mathbf{k},\theta}(m_1) < S_{\mathbf{q},\mathbf{k},\theta}(m_2)).$$

The difference between the two RoPE products,

$$D = S(m_1) - S(m_2),$$

may be separately viewed as the difference of two independent normal variables following Section B:

$$\widetilde{D} = \widetilde{S}_{M/2} - \widetilde{S}_M \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2),$$

where

$$\begin{aligned}\mu_1 - \mu_2 &= \sum_{n=\lambda(M/2)}^{\lambda(M)} a_n \cos\phi_n \leq \sum_{n=\lambda(M/2)}^{\lambda(M)} a_n, \\ \sigma_1^2 + \sigma_2^2 &= 1/2 \sum_{n=\lambda(M/2)}^{\lambda(M)} a_n^2 + \sum_{n=0}^{\lambda(M/2)} a_n^2.\end{aligned}$$

and the probability of position inversion should be

$$Pr(\widetilde{D} < 0) = \Phi\left(-\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),$$

where Φ is the cumulative distribution function of the standard normal distribution.

Assumption of regular rotary amplitudes To more accurately estimate the probability, we need the actual values of a_n and ϕ_n, which are dependent on $\mathbf{q}, \mathbf{k}$. However, if we assume that $\{a_n\}$ are regular enough, [8] i.e. not a single a_n dominates, and

$$\sum_n a_n/h \approx \sqrt{\sum_n a_n^2/h} \approx a_{\max} > 0,$$

then we may derive an estimation of the probability of position inversion.

Theorem 5. *If* $\sum_n a_n/h \approx \sqrt{\sum_n a_n^2/h} \approx a_{\max} > 0$*, then*

$$Pr(\widetilde{S}_{M/2} - \widetilde{S}_M < 0) \geq \Phi\left(-\log 2\sqrt{\frac{h}{\log B(\log M - 0.5\log 2)}}\right). \tag{5}$$

[8] see §Limitations for justification of this assumption.

Proof. When M is so large that $\lambda(M) = \lambda(M/2) = h$, then $\mu_1 = \mu_2$. In this case, $\widetilde{D} \sim N(0, \sigma_1^2 + \sigma_2^2)$, and the probability $Pr = 0.5$. Since $-\log 2\sqrt{\frac{h}{\log B(\log M - 0.5\log 2)}} < 0$, the right hand side of Eq. (5) is less than 0.5 and therefore less than the questioned probability.

When $M \leq B$, we have $h \geq \lambda(M) > \lambda(M/2)$,

$$\begin{aligned}\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \lesssim & \frac{\lambda(M) - \lambda(M/2)}{\sqrt{\lambda(M)/2 + \lambda(M/2)/2}}. \\ \leq & \frac{h\frac{\log 2}{\log B}}{\sqrt{h\frac{2\log M - \log 2}{2\log B}}} \\ = & \log 2\sqrt{\frac{h}{\log B(\log M - 0.5\log 2)}}.\end{aligned}$$

When $h = \lambda(M) > \lambda(M/2)$, we have

$$\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \lesssim \frac{h - \lambda(M/2)}{\sqrt{h/2 + \lambda(M/2)/2}}.$$

Let $x = \lambda(M/2) \in [h(1 - \log_B 2), h]$, and $f(x) = \frac{h-x}{\sqrt{h/2+x/2}}$. Since f is a continuous function that decreases with x, it reaches its maximum at $x = h(1 - \log_B 2)$, that is when $M = B$. Then, $f(x) = \log 2\sqrt{\frac{h}{\log B(\log B - 0.5\log 2)}} = \log 2\sqrt{\frac{h}{\log B(\log M - 0.5\log 2)}}$.

□

This gives a lowerbound for the probability of position inversion. As $\log M$ or $\log B$ increases, $\frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2}}$ decreases, and the probability of position inversion increases. As shown in the proof, when $M \to \Theta(B)$, this probability goes to $1/2$. On the other hand, if B increases, this probability also increases; if we apply a threshold probability α, and constrain $Pr < \alpha$, then as B increases, the maximum possible M decreases. See Table 3.

Table 3: Smallest M for selected B such that the probability of position inversion is no less than 0.3, when $h = 64$.

B	10^4	10^5	10^6	10^7	10^8
M	2.65×10^5	23361	4630	1457	612

C.2 Position Aliasing

Definition C.2. Let $S(m)$ be the RoPE product of $\mathbf{q}, \mathbf{k}$ at distance m, and $\hat{S}(m) = \hat{S}_{\texttt{dtype}}(m)$ be the numerical result of $S(m)$ using a certain datatype `dtype`. If for $m_1 \neq m_2$, $\hat{S}(m_1) = \hat{S}(m_2)$, then the pair (m_1, m_2) exhibits **position aliasing**.

This is not to be confused with the concept of "aliasing" in signal reconstruction, where a high frequency signal beyond the Nyquist Limit is mistaken as a lower one (Liu, 2026). This type of aliasing happens when $M > \Theta(B)$, as pointed out in Section 2.1.

For any integer $M > 0$ and error $\varepsilon > 0$, which is related to the datatype resolution, if we independently take two random distances $m_1, m_2 < M$, what is the probability such that $|S(m_1) - S(m_2)| < \varepsilon$?

Since m_1 and m_2 are independent, we can see the two sums as two i.i.d. normal variables, i.e.

$$\widetilde{S}_{m_1}, \widetilde{S}_{m_2} \overset{i.i.d.}{\sim} N(\mu, \sigma^2).$$

Then the difference $\widetilde{D}$ is also a normal variable:

$$\widetilde{D} = \widetilde{S}_{m_1} - \widetilde{S}_{m_2} \sim N(0, 2\sigma^2).$$

We have

$$\begin{aligned} Pr(|\widetilde{D}| < \varepsilon) =& \Phi(\frac{\varepsilon}{\sqrt{2}\sigma}) - \Phi(-\frac{\varepsilon}{\sqrt{2}\sigma}) \\ =& 2\Phi(\frac{\varepsilon}{\sqrt{2}\sigma}) - 1. \end{aligned}$$

In reality, ε is related to the computational precision. Now let us calculate the resolution limit of numerical operations involved.

Suppose the floating data type has f explicit fraction bits — for a number x in the normal range, this means a resolution of $\varepsilon_0(x) = \Theta(2^{-f}x)$. For BF16, FP16, FP32, FP64, $f = 7, 10, 23, 52$ respectively (Henry et al., 2019).

For two dot products to be considered "different", they must have different attention scores, i.e. different values after the normalized softmax. Let s be a positive integer. Let $\mathbf{x} \in \mathbb{R}^s$, and $\mathbf{y} = \text{softmax}(\mathbf{x}/\sqrt{d})$. For some $i \neq j$, we need to know the minimum value of $|x_i - x_j|$ such that $|y_i - y_j| < \varepsilon_0(y)$.

When $y_i = y_j = y$, disturb $x_i - x_j$. Then

$$\frac{\partial(y_i - y_j)}{\partial(x_i - x_j)} = \frac{y}{\sqrt{d}}.$$

This means that to ensure $y_i \neq y_j$,

$$|x_i - x_j| \geq \frac{\sqrt{d}\varepsilon_0(y)}{y} = \Theta(2^{-f}\sqrt{d}).$$

This is the effective resolution for the dot product, i.e. $\varepsilon_s = \Theta(2^{-f}\sqrt{d})$, from the restriction of softmax.

Also, the dot product itself has the resolution limit of $\varepsilon_d = \Theta(2^{-f}\sum_n a_n)$. Therefore the final resolution constraint is

$$\varepsilon = \Theta\left(2^{-f}\max\left(\sqrt{d}, \sum_n a_n\right)\right).$$

In our hypothesis that no single a dominates,

$$\sqrt{\sum_{n<\lambda(M)} a_n^2/|\lambda(M)|} \sim \sum_{n<h} a_n/h = \Theta(a_{\max}).$$

Therefore,

$$\varepsilon/\sigma = \Theta(2^{-f}\max(\sqrt{d}, ha_{\max})/\sigma) = \Theta(2^{-f}\max\left(\sqrt{d}/(\sqrt{\lambda(M)}a_{\max}), h/\sqrt{\lambda(M)}\right)).$$

Consider a simplified case where $a_n = 1$. For a RoPE base of 10,000, if we use FP16, the probability of having the same positional value is 5.6‰ at 32k context length. If the RoPE base increases to 100,000, then the probability becomes 6.5‰. For a random **pair** of distances, this is the probability of the pair having the exact same attention score. This may not seem like a large number, but if we consider that there are $O(M^2)$ pairs of position, this leads to an astonishing 3.5 million pairs of possible position aliasing.

Theorem 6. *Uniformly and randomly choose distance $m_1 \in [0, M)$. Let $p(M)$ denote the probability that there exists a different distance $m_2 \in [0, M)/\{m_1\}$ such that $|S(m_1) - S(m_2)| < \varepsilon$, where $\varepsilon = \Theta\left(2^{-f} \max\left(\sqrt{d}, \sum_n a_n\right)\right)$ is the absolute resolution limit for the datatype with f explicit fraction bits. We have $p(M) \geq 1 - (1 - E)^{M-1}$ and $\lim_{M\to+\infty} p(M) = 1$.*

Proof. For a random m_1 in a given context length M, the target probability is

$$p(M) = 1 - (1 - Pr(|\widetilde{D}| < \varepsilon))^{M-1}.$$

Since

$$Pr(|\widetilde{D}| < \varepsilon) \geq 2\Phi\left(2^{-f} h / \sqrt{\lambda(M)}\right) - 1 \geq 2\Phi(2^{-f}\sqrt{h}) - 1,$$

let $E = 2\Phi(2^{-f}\sqrt{h}) - 1$. Then

$$p(M) \geq 1 - (1 - E)^{M-1}.$$

Since $p(M) \leq 1$, we have

$$\lim_{M\to+\infty} p(M) = 1.$$

□

C.3 Token Inversion

For the token identification objective, we want to consider how relevant the key token (corresponding to the given $\mathbf{k}$) is to the query token (corresponding to the given $\mathbf{q}$).

Definition C.3. For a query vector $\mathbf{q}$ and two key vectors $\mathbf{k}_1, \mathbf{k}_2$, if $S_{\mathbf{q},\mathbf{k}_1}(0) < S_{\mathbf{q},\mathbf{k}_2}(0)$, but for some $m > 0$, $S_{\mathbf{q},\mathbf{k}_1}(m) > S_{\mathbf{q},\mathbf{k}_2}(m)$, then a **token inversion** occurs at m.

For the following analysis, we apply an important simplification to the definition. Since for the same $(\mathbf{q}, \mathbf{k})$ pair, the theoretical upper bound of the RoPE product, denoted by $S_{\max}$, is

$$\sum_{n=0}^{h-1} a_j \cos(m\theta^n + \phi_n) \leq \sum_{n=0}^{h-1} a_n \equiv S_{\max},$$

for the pair $(\mathbf{q}, \mathbf{k})$, we may introduce a hypothetic key token, called the "prime" token, whose key vector is denoted by $\mathbf{k}'$, that reaches this RoPE product at $m = 0$. That is to say, The pair $(\mathbf{q}, \mathbf{k}')$ shares the same set of amplitudes, a_j, as $(\mathbf{q}, \mathbf{k})$, with all initial phase biases $\phi_j = 0$. Now, if we randomly select $0 \leq m < M$, we wish to see the probability that this "prime" key token has a smaller RoPE product than a key token at distance m.

Formally, for a given pair $(\mathbf{q}, \mathbf{k})$, let the "prime" key vector be $\mathbf{k}'$. Randomly and uniformly select integer m between $[0, M)$. We assume that $\{\phi_n\}$ are not all 0, since otherwise $\mathbf{k} = \mathbf{k}'$. We also assume $a_{\max} > 0$, otherwise either $\mathbf{q}$ or $\mathbf{k}$ is zero.

Using the conclusion in Section B, the difference is

Table 4: Prob. of positional aliasing for a single pair when $\mathbf{a} = \mathbf{1}$.

M	B	Prob. at bf16	fp16	fp32
		Expectations		
1024	10000	0.057 3e+04 ± 2e+02	0.0071 3.7e+03 ± 6e+01	8.7e-07 0.46 ± 0.7
	100000	0.064 3.3e+04 ± 2e+02	0.008 4.2e+03 ± 6e+01	9.7e-07 0.51 ± 0.7
	1000000	0.069 3.6e+04 ± 2e+02	0.0087 4.5e+03 ± 7e+01	1.1e-06 0.56 ± 0.7
4096	10000	0.052 4.4e+05 ± 6e+02	0.0065 5.5e+04 ± 2e+02	8e-07 6.7 ± 3e+00
	100000	0.058 4.9e+05 ± 7e+02	0.0073 6.1e+04 ± 2e+02	8.9e-07 7.4 ± 3e+00
	1000000	0.064 5.4e+05 ± 7e+02	0.008 6.7e+04 ± 3e+02	9.7e-07 8.2 ± 3e+00
16384	10000	0.048 6.5e+06 ± 2e+03	0.006 8.1e+05 ± 9e+02	7.4e-07 9.9e+01 ± 1e+01
	100000	0.054 7.3e+06 ± 3e+03	0.0068 9.1e+05 ± 1e+03	8.3e-07 1.1e+02 ± 1e+01
	1000000	0.059 8e+06 ± 3e+03	0.0074 1e+06 ± 1e+03	9.1e-07 1.2e+02 ± 1e+01
32768	10000	0.047 2.5e+07 ± 5e+03	0.0058 3.1e+06 ± 2e+03	7.1e-07 3.8e+02 ± 2e+01
	100000	0.052 2.8e+07 ± 5e+03	0.0065 3.5e+06 ± 2e+03	8e-07 4.3e+02 ± 2e+01
	1000000	0.057 3.1e+07 ± 5e+03	0.0071 3.8e+06 ± 2e+03	8.7e-07 4.7e+02 ± 2e+01
65536	10000	0.045 9.7e+07 ± 1e+04	0.0056 1.2e+07 ± 3e+03	6.9e-07 1.5e+03 ± 4e+01
	100000	0.051 1.1e+08 ± 1e+04	0.0063 1.4e+07 ± 4e+03	7.7e-07 1.7e+03 ± 4e+01
	1000000	0.055 1.2e+08 ± 1e+04	0.0069 1.5e+07 ± 4e+03	8.4e-07 1.8e+03 ± 4e+01

$$D = \sum_{n=0}^{h-1} a_n \cos(m\theta^n + \phi_n) - \sum_{n=0}^{h-1} a_n \cos(m\theta^n)$$
$$= -2 \sum_{n=0}^{h-1} a_n \sin(\phi_n/2) \sin(m\theta^n + \phi_n/2)$$

Let $A_n = -2a_n \sin(\phi_n/2)$, $\varphi_n = \phi_n/2 - \pi/2$. Then

$$D = \sum_{n=0}^{h-1} A_n \cos(m\theta^n + \varphi_n).$$

This form is exactly what is studied in Section B. Just like any RoPE product, D can also be seen as a normal variable $\widetilde{D} \sim N(\mu, \sigma)$.

Theorem 7. *The probability* $Pr\left(\widetilde{D} > 0\right)$ *satisfies*

$$Pr\left(\widetilde{D} > 0\right) = \Phi\left(\frac{\mu}{\sigma}\right), \tag{6}$$

where

$$\mu = \sum_{n \geq \lambda(M)} a_n(\cos\phi_n - 1),$$
$$\sigma = \sqrt{\sum_{n<\lambda(M)} a_n^2 \sin^2\phi_n}.$$

When $M = \Theta(B)$ *and* $\lambda(M) = h$, $Pr\left(\widetilde{D} > 0\right) = 1/2$.

Proof. μ and σ can be obtained by applying Eq. (3) and Eq. (4) on $\{A_n\}$ and $\{\varphi_n\}$. The probability then follows that for normal distribution.

When $\lambda(M) = h$, $\mu = 0$. Since we assume that $\{\phi_n\}$ are not all 0 and $\{a_n\}$ are not all 0, $\sigma \neq 0$, giving $Pr\left(\widetilde{D} > 0\right) = 1/2$. □

C.4 Token Aliasing

Definition C.4. For a query vector $\mathbf{q}$ and two key vectors $\mathbf{k}_1, \mathbf{k}_2$, if the numerical results of RoPE product under a certain datatype $\hat{S}_{\mathbf{q},\mathbf{k}_1|\texttt{dtype}}(m) = \hat{S}_{\mathbf{q},\mathbf{k}_2|\texttt{dtype}}(m)$ for some m, then a **token aliasing** occurs at m.

For the purpose of simplicity, we make an assumption that in a fixed transformer head, for all pairs $(\mathbf{q}, \mathbf{k})$, for all $0 \leq n < h$, the magnitude at the given dimension pair $a_n(q,k) = |(q_{2n}, q_{2n+1})||(k_{2n}, k_{2n+1})|$ stays the same. For real attention heads, we may make the assumption that for a fixed frequency component n and randomly chosen q, k pairs, a_n follows some normal distribution, and we may take its mean as a_n discussed here.

If for query vector $\mathbf{q}$ and key vectors $\mathbf{k}_1$ and $\mathbf{k}_2$,

$$S_1(m) = \langle \mathbf{q}, \mathbf{k}_1 \rangle_m = \sum_{n=0}^{h-1} a_n \cos(m\theta^n + \phi_{1,n})$$

and

$$S_2(m) = \langle \mathbf{q}, \mathbf{k}_2 \rangle_m = \sum_{n=0}^{h-1} a_n \cos(m\theta^n + \phi_{2,n})$$

are independent, then for $\varepsilon > 0$, the probability of token aliasing can be expressed as

$$Pr(|S_1(m) - S_2(m)| < \varepsilon).$$

The difference $D = S_1(m) - S_2(m)$ can be seen as a normal variable

$$\widetilde{D} \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

Theorem 8. *If* $\sum_n a_n/h \approx \sqrt{\sum_n a_n^2/h} \approx a_{\max}$, *then*

$$Pr(|\widetilde{D}| < \varepsilon) \approx 2\frac{2^{-f}h}{\sqrt{\lambda(m)}} pdf\left(\frac{h}{\sqrt{\lambda(m)}} - \sqrt{\lambda(m)}\right),$$

where pdf is the probability density function of standard normal distribution,

$$pdf(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Proof. Without loss of generality, assume $\mu_1 > \mu_2$. We have

$$\mu = \mu_1 - \mu_2 = \sum_{n \geq \lambda(m)} a_n(\cos\phi_{1,n} - \cos\phi_{2,n}) \leq 2 \sum_{n \geq \lambda(m)} a_n,$$

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 = 2\sigma_1^2 = \sum_{n < \lambda(m)} a_n^2.$$

So

$$Pr(|\widetilde{D}| < \varepsilon) = \Phi\left(\frac{\varepsilon - \mu}{\sigma}\right) - \Phi\left(\frac{-\varepsilon - \mu}{\sigma}\right).$$

Since no a_n dominates, using the same estimations for position aliasing,

$$\begin{aligned} \mu/\sigma =& \Theta(\frac{h}{\sqrt{\lambda(m)}} - \sqrt{\lambda(m)}), \\ \varepsilon/\sigma \geq& \Theta(\frac{2^{-f}h}{\sqrt{\lambda(m)}}), \\ \Phi\left(\frac{\varepsilon - \mu}{\sigma}\right) - \Phi\left(\frac{-\varepsilon - \mu}{\sigma}\right) \approx& 2\frac{2^{-f}h}{\sqrt{\lambda(m)}}\text{pdf}(\frac{h}{\sqrt{\lambda(m)}} - \sqrt{\lambda(m)}). \end{aligned}$$

□

When $\lambda(m) = h$, the probability converges to $\Theta(2^{1-f}\sqrt{h}/\sqrt{2\pi})$.

D Experiment Details

D.1 Case Study

We use Head 0, Layer 0 of Llama3.1-8B (Grattafiori et al., 2024) as the case study sample. However, our method is applicable to any head in any layer of any RoPE-based decoder transformer model.

Each failure mode features a query token and one to two key tokens. For each key token, we calculate its RoPE product with the query, $S(m)$, for every m in the context range $(0, M]$. We calculate $S(m)$ for a certain head using the following steps:

- we construct an input following the format <bos>[9] `[key] ... [key] [query]`, where the key token is repeated M times.
- We calculate the hidden states for the target layer. This involves a forward pass through the embedding layer and every transformer layer before the one which contains our selected head. This standard process can be accelerated by FlashAttention (Dao et al., 2022).
- For the selected head, we obtain the *un-normalized* attention score, i.e. with neither the $\sqrt{d}$ normalization nor softmax. We only calculate the final row of the attention score matrix, since we are interested in the RoPE products involving the query token.
- The result, $\boldsymbol{o}$, is an $(M + 2)$-d array starting at index 0. For $0 < m \leq M$, we have $S(m) = o_{M+1-m}$.

The occurrences of the four failure modes are then identified using the definitions C.1, C.2, C.3[10], C.4 in Section C.

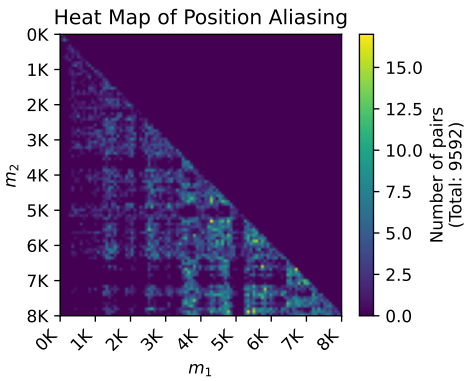
For larger models or heads not located in the first layer, GPUs with memory corresponding to the model size are recommended. However, for our case study, we only load the first layer and the whole case study is conducted using a lap-top grade CPU with 64GB memory.

[9]The <bos> token is added for the input sanity and is crucial if the target head is not from the first layer.

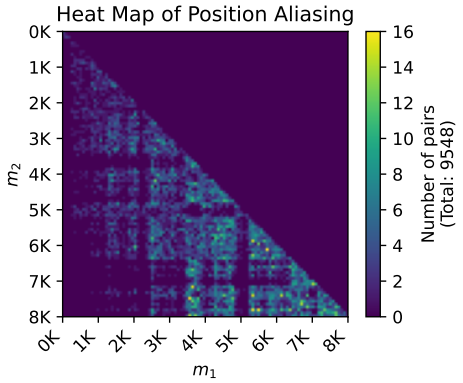
[10]For simplicity of implementation, we use $S(1)$ instead of $S(0)$ when identifying token inversion.

D.1.1 Aliasing Probs for FP16 in Case Study

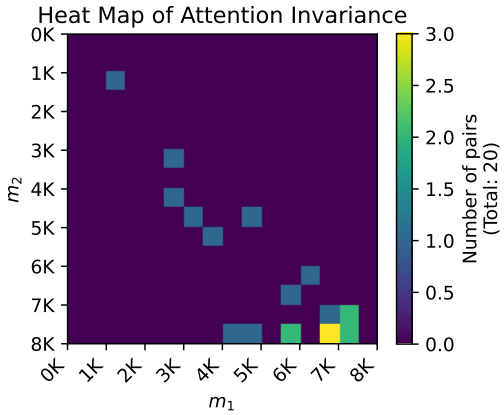
FP16 has 10 explicit fraction bits, and the aliasing probabilities are different from using BF16. For position aliasing, see Fig. 10.



(a) Distribution of position aliasing pairs for key "cat" and query "pet".



(b) Position aliasing pairs for key "dog" and query "pet".



(c) Attention invariance pairs for "cat", "dog" and "pet".

Figure 10: Heat maps of position aliasing and attention invariance pairs. FP16, Llama3.1-8B, Layer 0 Head 0. Pairs are grouped into a total of 200×200 bins for position aliasing, and 16×16 bins for attention invariance. 1K = 1024.

For token aliasing, see Fig. 11.

D.2 The Indexing Task

We randomly generate a list `arr`. The list only contains integers 0, 1, 2 and 3. We control the length of the list to be powers of 2, ranging from 4 to 4096, and use `f"{arr}"` to convert it into a string. The model is required to answer the value of the given index `arr[i]`. Different models may not tokenize the input or apply the chat template the same way, but the correspondence is roughly 3 tokens per element, and we only evaluate model performance based on token count. For each length, we test the models on 10 randomly generated lists. For each list, we generate 10 independent query sessions with random indices. For each list, we report the average input tokens and mean accuracy across the 10 query sessions. We aggregate the results for different lists by also reporting the mean standard deviation of accuracy.

We use the following prompt: `arr = {arr}\nGiven the above array, don't think and directly answer the corresponding value concisely: arr[{key}] =`

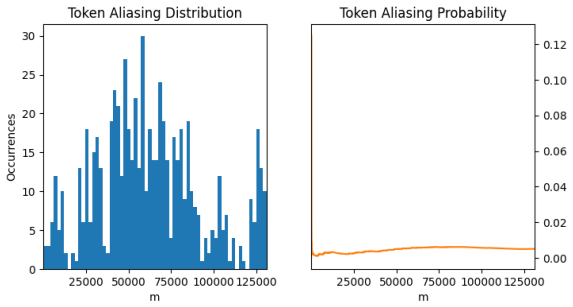


Figure 11: Distribution and probability of token aliasing for keys "cat" and "dog" and query "pet". Llama3.1-8B, Layer 0 Head 0. The probability converges to what is close to the estimated probability of 0.006.

We use the following models: smaller models (less than 10 B), including Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Mistral AI Team, 2024), Qwen3-8B (Yang et al., 2025a); larger models (more than 100B), including DeepSeek-V3.1 (DeepSeek-AI et al., 2025), Kimi-K2.5 (Team et al., 2026a), gpt-oss-120b (OpenAI et al., 2025).

For models less than 10B, we use the HuggingFace transformers implementation (Wolf et al., 2020). These inferences can be done on a GPU of more than 40GB. For models larger than 100B, we use the TogetherAI API (Together AI, n.d.). We disable reasoning mode where possible, and prompt the model to answer directly, but do not limit the generation length. We retrieve the last number from the model generation.

E RoPE in Real Models

As suggested in the experiment section §5, the multiple heads and layers in a real model may only offer redundancies that provide limited protection. We provide the following discussions that are open-ended and intuitive rather than evidence-based, which may serve as preliminary insights for future analyses using more realistic models:

First, redundancy across attention heads is limited, since the heads are highly specialized and sparsely activated. Prior work (Kahardipraja et al., 2025; Wu et al., 2024; Lin et al., 2026) has shown that different heads are often responsible for text dependencies of different ranges, and that only a small number of heads are active at the same time. As an example, consider a long input with a 50% probability of position inversion. Even if 8 out of 32 heads are actively retrieving, a proportion larger than what is commonly observed (Wu et al., 2024), once every 250 tokens, *all* retrieval heads will simultaneously exhibit position inversion.

Second, redundancy across layers is largely limited due to the residue connection. Errors introduced by earlier layers carry on to later ones connected in series. Improved architectures like Team et al. (2026b) apply residue connections in parallel. This aggregation strategy potentially reduces the accumulative error. Even in such cases, errors in the final layer never get the chance to be calibrated. An attention invariance failure in the final layer can directly lead to a wrong output token.

Finally, the underlying mechanism applies beyond the number of heads and layers. As context increases, an increasing oscillation leads to less positional uniqueness, and an increasing decay reduces the value of RoPE product and compresses token-wise difference. This means that the position and token identity of faraway text are increasingly likely to be poorly distinguished, leading to a weak or erroneous contribution to the context-aware representation.

F Limitations

Ambiguity of the frequency threshold It may be noticed in the illustrations of normal approximation of the RoPE product, that the real distribution is slightly skewed compared to our approximation. We provide only a rough threshold $\lambda(M) = \Theta(h \log_B M)$, following the common practice of divid-

ing the entire frequency domain into two categories of high and low. Strictly speaking, the threshold frequencies where $n = \Theta(h \log_B M)$ should be considered separately: their rotations are around 1 complete circle. We mostly ignore these terms, since there are only $O(h/\log B)$ of them, usually not exceeding 10. We believe that ignoring them only affects the precision of approximation, not the fundamental conclusions.

Assumption of regular rotary amplitudes We assume in this paper that the amplitudes of the RoPE products, $\{a_n\}$, are relatively uniform (no terms dominate). This is for simplicity in calculation. In real attention heads, we often observe several dimensions that have amplitudes significantly larger than others. Although we did not cover this in the paper, these dominating terms reduce the effects of other components; this heuristically works as if reducing the total number of dimensions and simplifying the oscillation patterns. The effective context length limit shortens to the maximum wavelength with a non-negligible amplitude. Smaller amplitudes for high frequency terms reduce oscillation; this may affect position identification for close position pairs, which is discussed in Liu (2026). In short, we assume that non-regular amplitudes are suboptimal, shortening the effective context length in the sense of the failure modes we discuss in this paper. We hope that future studies can provide more accurate mathematical models that may strictly remove this assumption.

Real models We do not theoretically analyze the multi-layer multi-head attention, or conduct experiments that categorize the failure modes, since real models are significantly more complicated, and one actual failure can be caused by mixed factors. Our final experiment merely serves as a demonstration that real models, regardless of their size, still face either position or token confusion, which we analyzed for a single head, and must sacrifice position confusion for better token identification. We do not intend to propose any new benchmarks or metrics, or make the task any realistic. While our experiments on real models do not directly prove that the failures are actually caused by the RoPE confusions, they indicate that the redundancy provided by multiple heads and layers may have limited protection from the long-context decay. In Section E, we provide discussions that may serve as a preliminary insight for future investigations on how the failures accumulate with heads and propagate across layers.

RoPE scaling While we show in Section B.1 that certain RoPE scaling methods may be reduced to standard RoPE, we do not provide a more detailed analysis for any specific variant. However, we use Llama-3.1 in our case study, which uses RoPE scaling, to illustrate that the scaling does not fundamentally resolve the problem, if it helps at all.

G Related Works

Long Context Models State-of-the-art language models are delivered with increasingly larger context window limits, some of them well beyond 1 million tokens (citations like Comanici et al. (2025); Meta (2025); OpenAI (2025); Magic (2024)). To deliver and utilize long-context models, apart from improvements in data curation (Fu et al. (2024b)), training optimization (Dao et al. (2022); Li et al. (2023); Liu et al. (2023a)), and efficient deployment (Xiao et al. (2023)), one of the main strategies is to adjust the value of RoPE base, which in the original paper is 10,000, "the worst base value" (Liu et al. (2023b)).

A War of Increasing RoPE Base Rotary Positional Embedding (RoPE, Su et al. (2021)) exhibits a decaying effect on attention scores of distant tokens. It is purposefully designed this way to model natural language with decreased dependency over longer distance (Su et al. (2021)). However, it is widely believed that this decay at least partially limits long-context performance (Tworkowski et al. (2023); Xu et al. (2024); Zhong et al. (2024); Miranda et al. (2024)). Therefore, attempts to extend context length usually feature increased values of RoPE base, which leads to slower decay (Peng et al. (2024); Gao et al. (2025)). As a variant, some works remove RoPE at least partially from attention calculation (Wang et al., 2024; Yang et al., 2025b), or only apply RoPE to certain dimensions (Javaheripi et al., 2023; DeepSeek-AI et al., 2026), which, according to our analysis, is equivalent to applying an infinitely large RoPE base. However, although the resulting long-context models excel in retrieval-based tasks such as Needle-in-a-Haystack (Kamradt (2023)), they still suffer a substantial performance drop on tasks that involve reasoning, variable tracking or other long-term

dependency, well within their context limit (Du et al. (2025); Hsieh et al. (2024); Kuratov et al. (2024); Liu et al. (2024)).

RoPE as a Function of Distance Multiple works have analyses on the dot product after applying RoPE (the RoPE product) as a function w.r.t. distance between two tokens. Works like Jonasson (2025); Liu et al. (2023b); Miranda et al. (2024) identify high and low frequencies based on whether a rotary component completes a full circle. Works that mainly focus on the decaying nature of RoPE, and its effect on the decreased attention score assigned to distant tokens (Xu et al. (2024); Xiong et al. (2024); Chen et al. (2024)), lead to suggested lower bounds of the RoPE base for certain context lengths (Xu et al. (2024); Peng et al. (2024)). However, these only focus on one side of the full picture: there are few systematic discussions about the oscillation in the waveform affected by the high-frequency terms, and how this can lead to upper bounds for the RoPE base selection. As a function of distance, the mathematical properties of the RoPE product are still little studied in depth, let alone their practical implications.

The Long-Context Dilemma Liu (2026) is among the first to study the oscillation using signal processing techniques, introducing a theoretical upper-bound where the numeric precision and Nyquist Limit are involved. Liu (2026) shows that as long as we must use transformer with RoPE to process long data, we must face some sort of uncertainty and are forced to choose between positional and token-wise accuracy. It is natural to think that transformers with RoPE have limited capacity to deal with long-context data. One must doubt whether the current ways of utilizing long-context models are the most adequate: are the reasoning models that rely on long chain of thoughts really capable of being responsible to their own thought process? Can chat models really recall distant history? Is it really worth it to train models with context lengths of tens of millions of tokens just to pass the Needle in a Haystack test? It is possible that to make language models really process large volumes of text, we need some other sorts of model structures, or strategies to keep the current models within their effective context range.