RoPE at the End of Its Rope? Theory, Diagnosis, and Mitigation of Long-Context Failures

Anonymous authors
Paper under double-blind review

Abstract

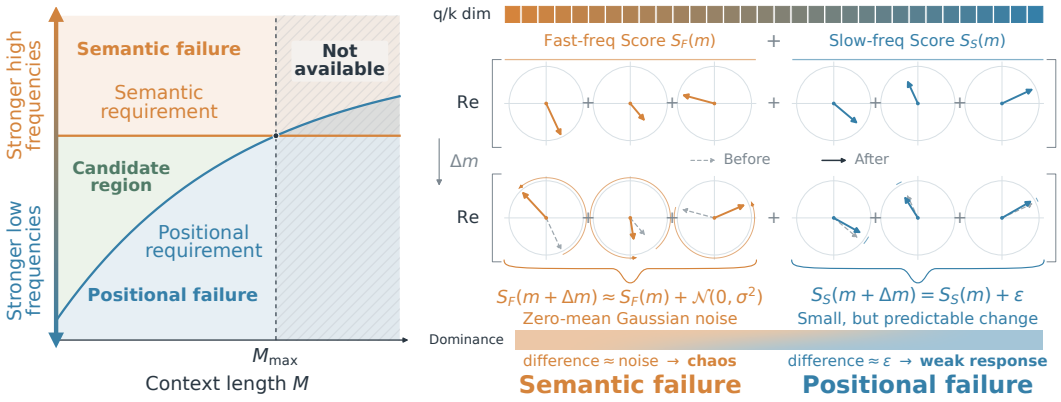
Long-context failures of RoPE-based language models can arise from RoPE's intrinsic tradeoff between maintaining stable token preferences and distinguishing nearby positions. Determining which weakness to address, and how, requires a more precise characterization of RoPE's behavior in trained models across context lengths. We address a key limitation of prior theory by allowing unequal query–key scales across RoPE frequencies, which aligns well with practical empirical observations. Our theory makes both vulnerabilities measurable for individual heads and inputs, and quantifies how high-frequency components support positional sensitivity while potentially disrupting semantic stability. We also derive **a theoretical context-length bound** beyond which, under specified conditions, a fixed attention-score comparison cannot jointly avoid *semantic reversal* and *positional insensitivity*. Guided by our fresh theoretical insights, we introduce **RoPE Profiler**, a *lightweight, plug-and-play diagnostic toolkit* that augments existing evaluations with zero additional forward passes by reusing cached query and key activations. Reusing activations collected during evaluation, the toolkit incurs little overhead. It supplements standard benchmark scores with two diagnostic scores that reveal semantic and positional weaknesses and help users prioritize which aspect to address. Crucially, our evaluations across 49 long-context task settings reveal a distinct pattern where reasoning tasks predominantly suffer from semantic reversal, whereas retrieval tasks are primarily vulnerable to positional insensitivity. Guided by our theory and diagnostic profiles, targeted high-frequency rescaling achieves immediate gains without additional training, improving task accuracy by up to 20 percentage points on Qwen3-8B and 25 percentage points on Llama-3.1-8B-Instruct.

1 Introduction

Long-context large language models (LLMs) open new possibilities for tackling complex, long-horizon problems, such as agentic software development and scientific discovery (Young, 2025; Zhao et al., 2026). Rotary position embedding (RoPE; Su et al., 2024) and its variants are widely used in these models (Dubey et al., 2024; Yang et al., 2025a; Gemma Team, 2025; Yang et al., 2024; Liu et al., 2026) and have consequently become a prominent target for improving long-context performance (Chen et al., 2023b; Peng et al., 2024; Ding et al., 2024; Barbero et al., 2025; Chen et al., 2025; Zhang et al., 2024; Wang et al., 2026; bloc97, 2023; An et al., 2025; Yang et al., 2025b; Xiong et al., 2025; Lu et al., 2025). Recent theoretical and empirical studies highlight a tension between preserving content preferences and distinguishing token positions as context length grows (Barbero et al., 2025; Du et al., 2026; Urrutia et al., 2026; Liu, 2026; Xu et al., 2024; Jerad et al., 2026). Practical progress based on these findings requires a more precise characterization of RoPE's tradeoff under assumptions that better reflect learned representations, together with diagnostics that provide actionable insights for targeted intervention. We advance both the theoretical and empirical fronts.

Our theory quantifies two fundamental failure modes of RoPE-based single-head attention in a single layer and establishes their quantitative tradeoff across expanding context scales (Figure 1a). As illustrated in Figure 1b, RoPE rotates coordinate pairs of query and key activations at geometrically spaced angular frequencies across relative token distances. These frequency channels, which the attention mechanism aggregates, have different functions: rapidly rotating components produce fluctuations whose mean is close to zero and whose variance we can bound analytically with a

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.



(a) Semantic-positional tradeoff **(b) Rotary frequency dynamics**

Figure 1: **The fundamental semantic-positional tradeoff and rotary frequency dynamics.** (a) The candidate region for semantic stability and adjacent positional sensitivity shrinks with context length. The orange and blue curves represent the semantic upper limit and positional lower limit on the high-frequency norm share. The green region contains candidate values satisfying both requirements; beyond $M_{\max}$, no such value remains within the calibrated family. (b) RoPE rotates coordinate pairs of query and key vectors by frequency-dependent angles across relative token distances. Rapidly rotating components, when aggregated across channels, form unpredictable Gaussian noise that destabilizes token preferences. Slowly rotating components remain predictable but produce marginal score changes that fail to resolve adjacent positions. Consequently, the dominance of different frequency bands triggers distinct failure modes. Precise formulations are given in §3.2 and Appendix A.5.

Gaussian approximation (Du et al., 2026; Salem & Zygmund, 1947; Aistleitner et al., 2024), whereas slowly rotating components are insensitive to adjacent positions and contribute little change to the attention score (Liu et al., 2024b; Hong et al., 2024). Consequently, these different functionalities triggers two fundamental failures: *semantic reversal* and *positional insensitivity* (§2). In semantic reversal, changing position alone reverses a query's preference between keys, potentially redirecting attention toward distractors (Bansal et al., 2026; Barbero et al., 2025). In positional insensitivity, a one-token shift changes the attention score too little to distinguish adjacent positions, which can hurt tasks that require retrieving elements immediately before or after a target, as shown in frequency-scaling based length-extension techniques (Peng et al., 2024; bloc97, 2023; Wu et al., 2026). As context length grows, Figure 1a illustrates how preserving adjacent positional sensitivity imposes a higher minimum high-frequency norm share, raising the calibrated lower bound on reversal risk and shrinking the candidate region. Under the stated calibration and parameter conditions, this tradeoff yields a context-length bound $M_{\max}$ beyond which a given attention-score margin can no longer avoid both failures within the calibrated windows (§3). Crucially, our finite-window bounds accommodate arbitrary fixed query and key vectors extracted from trained models, including unequal coefficient magnitudes across channels. By avoiding restrictive assumptions on token distributions or coordinate regularity across channels (such as in Xu et al. (2024); Barbero et al. (2025); Du et al. (2026); see also §I.1), our analysis establishes a realistic theoretical foundation for this spectral ceiling.

Our theory reveals constraints on mitigating both failures simultaneously. However, we also find that these constraints imply a principled basis for identifying which weakness to prioritize and how to address it for a given task. Building on our theoretical insights, we introduce **RoPE Profiler**, a lightweight diagnostic tool that seamlessly integrates into existing benchmarks (§4.1). RoPE Profiler reuses query and key activations cached during evaluation and incurs little additional computational overhead. RoPE Profiler turns the two theoretical failure criteria into practical diagnostics, supplementing standard benchmark scores with two metrics: the *Semantic Score* measures the stability of key preferences across positions, while the *Positional Score* measures sensitivity across adjacent positions. Our theory provides a quantitative basis for two training-free interventions: rescaling high-frequency components to adjust their relative strength and freezing their rotations to suppress position-induced fluctuations (Qiao et al., 2025; Barbero et al., 2025; Chen et al., 2025). Across 49 task configurations from nine benchmark sources (described in Appendix E), the evaluated models show broadly similar rankings of their relative semantic and positional weaknesses (§4.2). Guided by

these diagnostic profiles, our training-free interventions yield best observed accuracy gains of up to 20 percentage points on Qwen3-8B and 25 percentage points on Llama-3.1-8B-Instruct, demonstrating that spectral diagnosis provides effective, targeted guidance for long-context optimization (§4.2).

Although our theory aligns with Du et al. (2026)'s pessimistic assessment of RoPE's intrinsic long-context limitations, our precise spectral characterization reveals concrete opportunities for targeted improvements in practice. Our findings demonstrate that substantial optimization space remains simply by adjusting frequency allocations across individual attention heads. We believe further progress lies in incorporating RoPE Profiler's diagnostics into training and long-context adaptation. Beyond uniform frequency scaling, designing attention mechanisms that assign complementary roles to different heads can better balance content retrieval and positional reasoning (Shi et al., 2026; Wang et al., 2026; Xiong et al., 2025; Xiao et al., 2025; Tang et al., 2025; Wu et al., 2025a). In the longer term, fully overcoming these intrinsic constraints will require rethinking attention architectures and how positions are encoded in foundation models (Kimi Team, 2026; Press et al., 2022a; Kazemnejad et al., 2023; Yang et al., 2025b; Hua et al., 2025).

2 Two Failure Modes of RoPE: Semantic Reversal and Positional Insensitivity

A good positional embedding (PE) in transformers should achieve two objectives: **1) retain the attention's ability to distinguish different tokens:** for a given query, the attention mechanism distinguishes two keys by having different scores, and PEs should not get in the way of this feature; **2) distinguish different positions:** varying the score for the same key across locations informs the model of token positions (Su et al., 2024; Vaswani et al., 2017; Shaw et al., 2018; Press et al., 2022b). Unmasked attention with position-independent inputs is permutation-equivariant (Lee et al., 2019), although causal language models can learn positional information without explicit positional encodings (Haviv et al., 2022). RoPE encodes relative distance by rotating query and key coordinate pairs, making their attention score position-dependent.

2.1 Background

RoPE as a Sum of Rotated Complex Numbers. Let q and k denote the query and key vectors from an attention head. Their scaled dot product forms the pre-softmax attention score, which converts into weights for combining value vectors. RoPE (Su et al., 2024) injects position information by rotating coordinate pairs of queries and keys. Let p_q and p_k represent the token positions, and let $m = p_q - p_k$ denote their relative distance. For a head with h rotary coordinate pairs and base B, coordinate pair n rotates by angle $p\omega_n$ at position p, where $\omega_n = B^{-n/h}$ for $n = 0, \ldots, h-1$. While the base frequency ω_n decays smoothly with coordinate index n, individual components exhibit starkly disparate behaviors across long contexts.[1]

Following the complex-valued formulation of RoPE in Zhu et al. (2024), let $Q_n = q_{n,1} + iq_{n,2}$ and $K_n = k_{n,1} + ik_{n,2}$ denote the complex representations of each rotary pair, with $i^2 = -1$. With attention scale c_{att}, the rotary attention score simplifies to

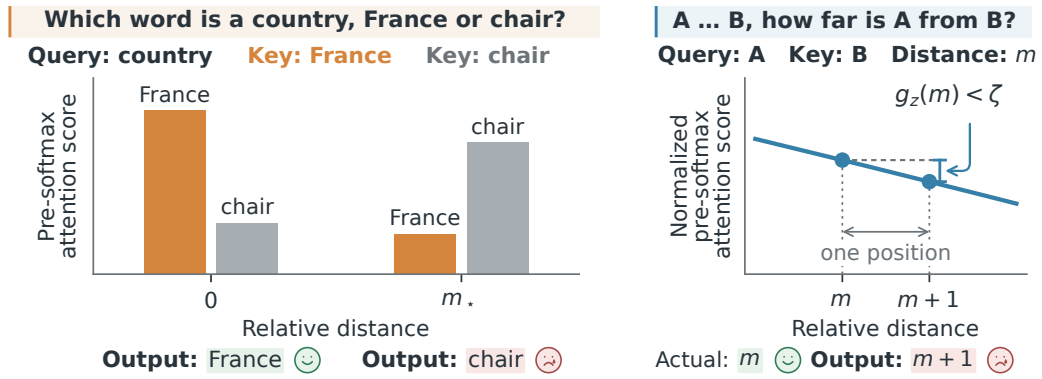
$$S_{q,k}(m) = \mathrm{Re} \sum_{n=0}^{h-1} z_n e^{im\omega_n}, \qquad z_n = c_{\text{att}} Q_n \overline{K_n}, \tag{1}$$

where Re takes the real part and the overline denotes complex conjugation. Here, token content determines the static amplitude coefficients z_n, while relative distance m modulates the dynamic rotation phases $m\omega_n$. We write $S(m)$ when the query and key are clear from the context.

The Janus Face of RoPE: Sustaining Content and Localizing Tokens. By applying frequency-dependent rotations across relative distances, RoPE enables attention to produce distinct scores

[1] For example, consider a typical setup with rotary dimension $2h = 128$, base $B = 500{,}000$ (as in Llama 3 Dubey et al. (2024)), and context length $L = 32{,}768$. Over this window, the fastest component with $\omega_0 = 1$ completes over $5{,}215$ full cycles, while the slowest component with $\omega_{h-1} \approx 2 \times 10^{-6}$ completes approximately 0.01 cycles, covering merely a tiny fraction of a single rotation.

according to both token content vectors and spatial positions. Extensive prior studies seek to demonstrate how RoPE successfully realizes these two core functions. These efforts span early decoupled representation designs, approximate additive decompositions and frequency-wise functional division in learned RoPE representations, examinations of locality and symmetry under role shuffling, and symbolic-versus-positional permutations that probe expressive capacity (Kitaev & Klein, 2018; Ke et al., 2021; Chen et al., 2023a; Han & Ji, 2025; Barbero et al., 2025; Urrutia et al., 2026; Oka et al., 2026a;b; Chen & Yan, 2024; Gu et al., 2026; Hong et al., 2024; Jin et al., 2025). In essence, these investigations explain *how RoPE succeeds*. However, analyzing its functional capability provides only an incomplete picture. The more critical question is how this dual responsibility becomes unbalanced as the context scale continues to expand.



(a) Semantic reversal **(b) Positional insensitivity**

Figure 2: **Illustrative task consequences of two RoPE failure modes.** (a) For query `country`, the scores of keys `France` and `chair` reverse order between virtual relative distances 0 and $m_\star$. As attention directs information routing toward higher scores, this reversal favors the distractor `chair` over the semantically relevant `France` at relative distance $m_\star$ and can contribute to an incorrect prediction. (b) For query `A` and key `B`, the vertical bracket marks a marginal normalized score gap between adjacent virtual relative distances, $g_z(m) < \zeta$. This leaves the model unable to distinguish between the two adjacent positions.

2.2 Formal Definitions of Semantic Reversal and Positional Insensitivity

RoPE gives attention a way to distinguish positions, but what is the cost? Examining this question through the lens of failure modes, Du et al. (2026) formalize four distinct vulnerabilities in RoPE-based attention. In this work, we focus on two core failure modes: *semantic reversal* and *positional insensitivity*. These two dimensions prove theoretically sufficient to uncover the fundamental context limits of RoPE in §3, while remaining practically informative for our diagnostic tool design in §4. Furthermore, our theoretical framework naturally extends to the other failure modes analyzed in prior literature, which we detail in Appendix B. Real-world tasks do not necessarily suffer from both vulnerabilities equally, and in §4.2 we illustrate how specific task families are governed by distinct failure modes. Below, we formalize the mathematical definitions of these two primary failure modes.

Semantic reversal (Different Tokens, Same Position). Motivated by Du et al. (2026), we examine the stability of attention preferences between distinct tokens across varying relative distances. Figure 2a illustrates an instance of semantic reversal. For a query and two keys, let $D(m) = S_+(m) - S_-(m)$ denote their ordered score margin. Without loss of generality, we assume $D(0) > 0$. A semantic reversal occurs at any relative distance where $D(m) < 0$. To capture overall behavior without being biased by individual positions, we evaluate how frequently semantic reversals occur across the entire context window of length M as the semantic reversal probability:

$$p_{\text{rev}}(D; M) := \frac{1}{M} \sum_{m=0}^{M-1} \mathbf{1}_{\{D(m)<0\}} = \mathbb{P}[D(\mathbf{m}) < 0], \qquad \mathbf{m} \sim \text{Unif}\{0, \ldots, M-1\}. \tag{2}$$

The probability evaluates the stability of the model's preference across the full sequence.

Positional insensitivity (Same Token, Different Positions). Motivated by Liu (2026); Sun et al. (2026), we evaluate how sensitively an attention score responds to one-token changes in relative distance while holding token representations fixed. For distant or semantically unrelated token pairs, resolving exact single-token offsets carries little operational value; however, in retrieval tasks that demand precise relative order, as illustrated in Figure 2b, distinguishing adjacent positions becomes essential to resolve local sequence structure (Wu et al., 2025b; Wang et al., 2024a).

An intuitive metric is to measure the raw score change $|S(m+1) - S(m)|$ across adjacent locations. However, this raw difference scales with representation magnitude, leaving the raw gap vulnerable to arbitrary norm variations across heads (Qi et al., 2026; Jin et al., 2025). To isolate intrinsic positional sensitivity from overall vector magnitude, we normalize the local score difference by the pair's coefficient norm $U(z) = \|z\|_2 = (\sum_n |z_n|^2)^{1/2}$. The resulting *normalized adjacent response* is

$$g_z(m) := \frac{|S(m+1) - S(m)|}{U(z)}, \qquad U(z) > 0, \quad 0 \le m \le M-2. \tag{3}$$

For a chosen response threshold $\zeta > 0$, *positional insensitivity* occurs when $g_z(m) < \zeta$, quantifying the failure to preserve local positional resolution.

3 A Theory of Long-Context Failures: The Tradeoff Between Semantic Stability and Positional Sensitivity

RoPE relies on a shared set of coordinate rotations to preserve content preferences and distinguish positions. In this section, we reveal a tradeoff between semantic stability and positional sensitivity in RoPE, both governed by the norm share of **high-frequency** components. Over long context windows, *an insufficient high-frequency contribution leaves attention blind to adjacent positions, whereas amplifying these fast oscillations triggers semantic reversal*. Below, we formalize high-frequency score variation under a finite-window framework, and then derive **a context-length limit** where avoiding both failure modes becomes mathematically impossible.

3.1 High-Frequency Norm Share and Score Variation

To analyze high-frequency fluctuations rigorously, we first specify an analytical moment tolerance $\varepsilon \in (0, 1/2)$ that governs the precision of finite-window moment control. Under the standard geometric grid $\omega_n = \rho^n$ with $\rho = B^{-1/h}$, demanding a tighter analytical tolerance requires a higher frequency cutoff, yielding the certified high-frequency band

$$H = \left\{ n \in \{0, \dots, h-1\} : M\omega_n \ge \frac{2C_{\mathrm{F}}}{(1-\rho)\varepsilon} \right\}, \tag{4}$$

where $C_{\mathrm{F}} > 0$ denotes the universal Fourier-frame bound established in Appendix C.2. This ε-dependent criterion guarantees that all coordinates in H complete sufficient oscillations over window M to suppress cross-frequency interference.

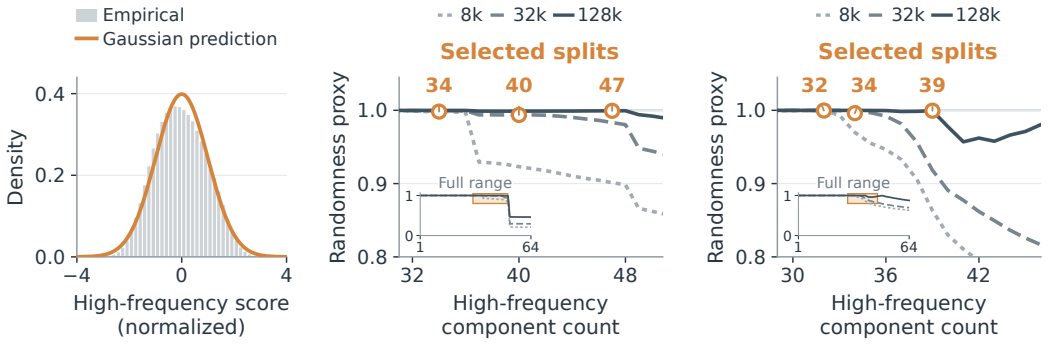
Proposition 1 (Analytical moment guarantees under finite windows). *For any tolerance $\varepsilon \in (0, 1/2)$ and certified window M where H is nonempty, the high-frequency attention score $S_H(m) := \Re \sum_{n \in H} z_n e^{im\omega_n}$ and its coefficient norm $U_H(z) := (\sum_{n \in H} |z_n|^2)^{1/2}$ over uniformly distributed relative distances $m \sim \mathrm{Unif}\{0, \dots, M-1\}$ satisfy*

$$|\mathbb{E}_m[S_H(m)]| \le \varepsilon U_H(z) \qquad \textit{and} \qquad \left| \mathrm{Var}_m(S_H(m)) - \frac{1}{2} U_H(z)^2 \right| \le \varepsilon U_H(z)^2. \tag{5}$$

Full proofs are deferred to § A.3. Following the statistical behavior observed by Du et al. (2026), which is an application of the lacunary Central Limit Theorem (Salem & Zygmund, 1947; Aistleitner et al., 2024), sums of rapidly oscillating components across relative distances behave as zero-mean Gaussian fluctuations, yielding the operational score approximation $S_H(m) \approx \mathcal{N}(0, U_H(z)^2/2)$.

Error-Bounded Spectral Partitioning. Proposition 1 links the frequency cutoff H directly to the analytical error tolerance ε, establishing precise control over finite-window moments. Demanding tighter precision forces slower components out of H, raising the frequency cutoff accordingly.

While prior analyses assume specific query and key distributions or amplitude regularity to derive bounds (Xu et al., 2024; Barbero et al., 2025; Du et al., 2026), our finite-window moment bounds hold for arbitrary fixed activations extracted from trained models, accommodating general phases and non-uniform coefficient magnitudes. Figure 3 illustrates the empirical behavior of our theoretically predicted partitioning across 8k, 32k, and 128k relative-distance windows.



(a) Qwen3-8B distribution **(b) Qwen3-8B split selection** **(c) Llama-3.1-8B split selection**

Figure 3: **(a)** Under our high-frequency split, the predicted score distribution closely matches the empirical distribution. **(b,c) Defining the high-frequency band.** The curves show how randomness changes as more components are included. Our theory-guided split selects the largest band satisfying the split criterion while retaining high randomness. The trends vary with context length, and larger M allows more high-frequency components, consistent with our theory. (See Appendix D.2.)

The Governing Norm Share r_H For context length M, we quantify the proportion of energy concentrated in fast oscillations through the high-frequency norm share

$$r_H(z; M) := \frac{U_H(z)}{U(z)} \in [0, 1], \qquad \text{where } U(z) := \|z\|_2 = \left(\sum_{n=0}^{h-1} |z_n|^2 \right)^{1/2}. \tag{6}$$

The parameter $r_H(z; M)$ serves as the central mathematical pivot connecting spectral allocations to downstream failure modes. Specifically, the two failure modes defined in §2.2 constrain this energy share in opposing directions. For an ordered score margin, the calibrated lower bound on semantic reversal probability is non-decreasing in r_H (Appendix A.5), while retaining adjacent positional sensitivity for the same margin imposes a context-dependent lower bound on r_H (Appendix A.4.1). This dual constraint establishes the foundation for the fundamental tradeoff derived in §3.2.

3.2 The Semantic and Positional Tradeoff Limits Context Length

Distinguishing adjacent positions requires attention scores to vary, while preserving token preferences demands that their relative ordering remains stable. High-frequency components govern both requirements. Suppressing these oscillations stabilizes token orderings at the expense of adjacent positional sensitivity. As the context window expands, slower frequencies enter the high-frequency band, leaving fewer static components to sustain local differences. Maintaining a target positional gap imposes a context-dependent lower bound on the high-frequency norm share, yielding a non-decreasing lower bound on reversal probability under the calibration conditions below.

Theorem 1 (Context-length limit on joint reliability (informal))**.** *For a fixed ordered score margin and reversal and adjacent-response tolerances δ and ζ satisfying the conditions in Appendix A.5, there exists a context-length bound $M_{\max}$ such that, for every calibrated window,*

$$M > M_{\max} \implies p_{\text{rev}}(M) > \delta \quad \textit{or} \quad \max_{0 \le m \le M-2} g(m) < \zeta, \tag{7}$$

where $p_{\text{rev}}(M)$ and $g(m)$ denote the reversal probability and normalized adjacent response of the same margin over relative distance m. Beyond $M_{\max}$, this margin cannot satisfy both requirements within the calibrated family.

Appendix A.5 gives the precise calibrated-window conditions, the construction of $M_{\max}$, and the formal version of Theorem 1. The construction balances the high-frequency share required for local sensitivity against the reversal probability it induces.

A RoPE-of-War between Content and Position (Diagnose What the Task Needs First). While existing literature identifies conceptual tensions in rotary attention regarding content selectivity and positional tracking (Barbero et al., 2025; Du et al., 2026; Urrutia et al., 2026), we formalize a quantitative tradeoff showing that semantic stability and positional sensitivity are mutually constrained under rotary representations. Theorem 1 gives a conditional bound on joint score-level reliability, determined by the frequency grid, the two reliability tolerances, and the calibration error. Within this setting, shifting spectral weight changes the balance between two requirements. The analysis suggests weakening high-frequency contributions to stabilize key preferences and strengthening them to distinguish adjacent positions, depending on the task. §4 operationalizes this principle by introducing **RoPE Profiler** to evaluate these competing vulnerabilities and guide targeted interventions.

4 Diagnosis and Targeted Mitigation of Long-Context Failures with RoPE Profiler

Building on our theoretical criteria, we introduce **RoPE Profiler** to turn spectral failure analysis into actionable diagnostics and targeted model adjustments. Moving beyond prior diagnostic frameworks that primarily identify internal vulnerabilities (Appendix I.4), our tool establishes a bridge between spectral state profiling and targeted training-free mitigation. Figure 4 outlines the five-stage pipeline. In §4.1, we formalize how cached activations yield the *Semantic Score* and *Positional Score*, evaluate relative failure susceptibility across benchmark settings, and derive theory-motivated interventions. In §4.2, we evaluate these diagnostic profiles and targeted modulations across diverse tasks and models, demonstrating substantial accuracy gains across a comprehensive long-context suite.

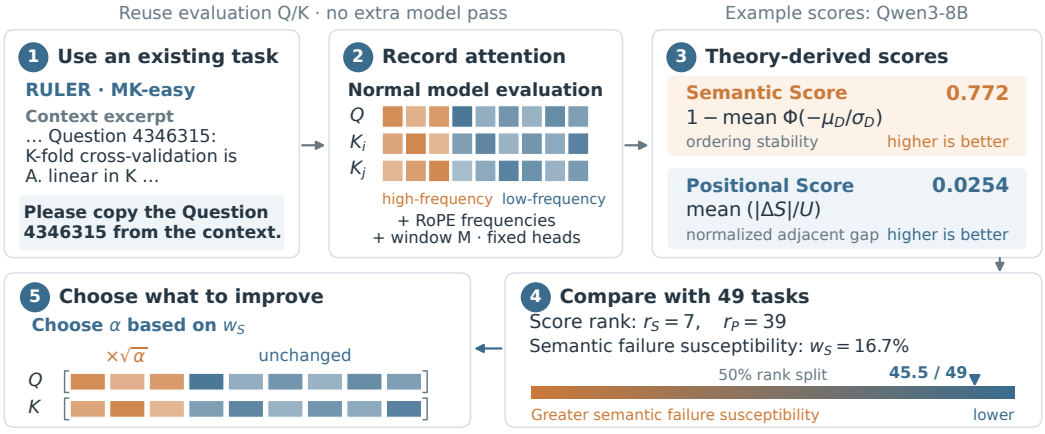


Figure 4: **Augmenting accuracy-based evaluation with failure diagnostics.** For a given model and the 49 reference task configurations (described in §E), we use query and key activations from standard evaluation to compute relative semantic failure susceptibility w_S. A higher w_S suggests greater susceptibility to semantic reversal relative to positional insensitivity within the reference set.

4.1 From Spectral Theory to Practical Profiling and Interventions

During the standard inference pass of benchmark evaluation, RoPE Profiler samples query and key tokens to cache their internal activations q, k across target attention heads (selection details and heatmaps in Appendix F) (Figure 4, Steps 1 and 2). Holding these content vectors fixed, we evaluate virtual attention scores by sweeping RoPE relative distances across the context window, incurring zero additional model inference passes.

Theoretical Scores from Cached States (Figure 4, Step 3). Our theoretical criteria translate directly into two continuous metrics. For local positional sensitivity, we evaluate query–key pairs

(q, k) by averaging the normalized adjacent response $g_{q,k}(m) = |S(m+1) - S(m)|/U(z)$ across relative distances m, defining the *Positional Score* as $\bar{g}$. For semantic stability, we construct triplets (q, k_1, k_2) with distinct keys and compute the finite-window probability P_{rev} that position shifts invert their reference ranking at distance zero, using our moment bounds and Gaussian approximation. The *Semantic Score* is defined as $1 - \bar{P}_{\text{rev}}$ to ensure that higher values are better for both metrics.

Relative Failure Susceptibility via Normalized Ranks (Figure 4, Step 4). Directly comparing raw scores across different tasks or failure modes is ineffective because absolute score magnitudes lack a shared scale, and meanwhile semantic and positional metrics operate on distinct numerical ranges. Therefore, we rank all evaluated tasks separately by each metric, yielding semantic rank r_S and positional rank r_P. We use the relative semantic failure susceptibility $w_S \in [0, 1]$ to show the normalized rank gap: a larger w_S indicates greater susceptibility to semantic reversal relative to positional insensitivity within the reference set.

Theory-Motivated Interventions (Figure 4, Step 5). The spectral tradeoff driven by the high-frequency norm share motivates a training-free intervention approach with two opposing directions on task-sensitive heads. Guided by task diagnostic profiles, we assign the upper half of w_S rankings to the semantic direction, attenuating high frequencies or freezing their rotations to suppress position-induced fluctuations (Barbero et al., 2025; Chen et al., 2025). The remaining settings follow the positional direction, using $\alpha > 1$ to amplify adjacent score resolution (Qiao et al., 2025; Chiang & Yogatama, 2025; Mikaeili et al., 2026). While w_S determines the intervention direction, the specific scaling scalar α is selected from candidate values. This strategy departs from conventional frequency manipulations (Hua et al., 2025; Xiong et al., 2025; Li et al., 2026; An et al., 2025) by dynamically assigning the intervention direction per task from its diagnostic profile and restricting modifications strictly to task-sensitive heads. We report the best observed outcome for each setting, while Appendix H documents complete sweeps across all candidate configurations.

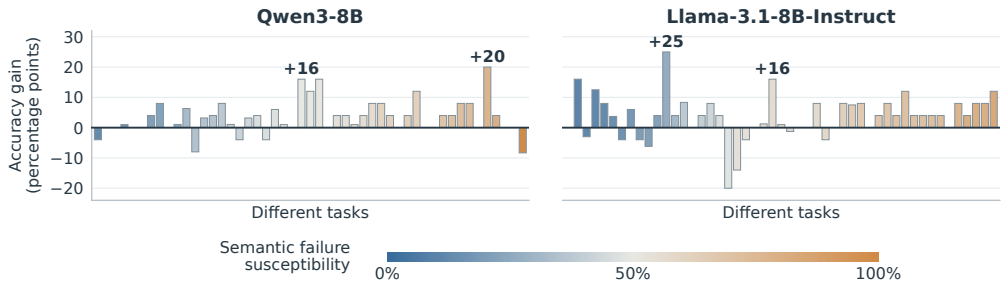


Figure 5: **Accuracy gains from our exploratory intervention search.** Bars show the best observed intervention in each setting's assigned semantic or positional direction minus the original baseline on identical examples within each setting. The baseline is excluded from the candidate maximum. Each model's 49 task/configuration settings (described in Appendix E) are ordered from left to right by increasing semantic failure susceptibility, from blue to orange.

4.2 Empirical Validation Across Diverse Long-Context Tasks

Targeted Intervention Gains Reveal Untapped Potential. Guided by diagnostic susceptibility profiles, our targeted training-free interventions achieve performance gains across a majority of evaluated tasks (63.3% of Qwen configurations and 65.3% of Llama configurations). As illustrated in Figure 5, accuracy gains can reach 20% on Qwen3-8B and 25% on Llama-3.1-8B-Instruct, showing the practical value of adjusting the spectral contributions identified by our theory.

RoPE Profiler Reveals Intrinsic Task Properties. Beyond intervention gains, Figure 6 compares semantic failure susceptibility across the 49 shared settings to inspect how failure patterns relate to underlying task structures. Across the the majority of tasks, the two models exhibit consistent susceptibility rankings with a Spearman correlation of $\rho = 0.658$, demonstrating that the diagnostic captures intrinsic task properties.

Specifically, *reasoning tasks cluster in the upper half where semantic reversal dominates, while multi-key retrieval tasks occupy the lower half governed by positional insensitivity*. In contrast, adjacent-element retrieval tasks deviates sharply between models, ranking in the top three for Qwen but between ranks 34 and 38 for Llama, underscoring that hybrid tasks coupling semantic matching with local order also reflect model-specific representation choices. These patterns align with previous observations on different task settings (Goldman et al., 2024; Vodrahalli et al., 2024; Kuratov et al., 2024; Modarressi et al., 2025; Hsieh et al., 2024a).

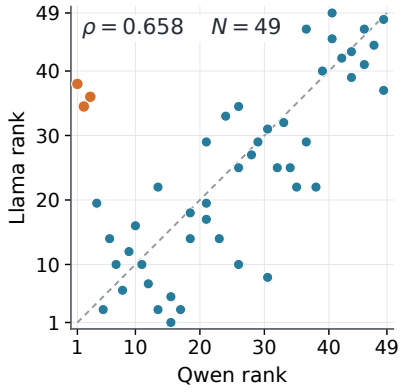


Figure 6: **Semantic failure susceptibility across tasks.** Points compare ranks of semantic failure susceptibility w_S in Qwen3-8B and Llama-3.1-8B-Instruct for the 49 task/configuration settings described in Appendix E. Most tasks show similar patterns across models, while three adjacent-element retrieval tasks differ.

5 Related Work

RoPE and context extension RoPE (Su et al., 2024) is widely used in open LLMs (Dubey et al., 2024; Yang et al., 2025a; Liu et al., 2026). Existing extension methods involve frequency rescaling, including NTK-aware scaling (bloc97, 2023), PI (Chen et al., 2023b), ABF (Xiong et al., 2024), YaRN (Peng et al., 2024), LongRoPE (Ding et al., 2024), and Resonance RoPE (Wang et al., 2024b). Others manipulate frequency components (Barbero et al., 2025; Hua et al., 2025; Chen et al., 2025; Li et al., 2026; Xiong et al., 2025), adjust scaling per head (Wang et al., 2026), or reindex positions (An et al., 2025; Zhang et al., 2024). These methods are task-agnostic, while we diagnose task-specific failures before deciding the intervention direction.

Theory of RoPE frequencies Existing RoPE theories derive base-dependent context bounds (Xu et al., 2024; Liu et al., 2024b; Liu, 2026) and roles of low-frequency dimensions (Hong et al., 2024; Jin et al., 2025). Prior works observe a tension between semantic and positional resolution (Barbero et al., 2025; Du et al., 2026; Urrutia et al., 2026). These analyses assume specific distribution properties of vector space or amplitude regularity, whereas our bounds hold for real-world activations without additional assumptions.

Internal diagnostics and specialized heads Retrieval heads (Wu et al., 2025a), attention sinks (Xiao et al., 2024), position-wise features (Dong et al., 2024), positional attention bias (Hsieh et al., 2024b), and other internal signals (Tan et al., 2026) explain long-context behavior inside the model. Our RoPE Profiler also utilizes the cached activations but both identify theoretic vulnerabilities and diagnose with task-specific failure susceptibility. For further discussion, see Appendix I.

6 Conclusion

In this paper, we have studied long-context failures of RoPE through the tradeoff between semantic stability and positional sensitivity. High-frequency components support fine positional distinctions while introducing fluctuations that can destabilize token preferences. Our finite-window analysis accommodates unequal coefficient magnitudes across frequencies and, under the stated calibration conditions, yields a context-length bound on jointly avoiding semantic reversal and positional insensitivity. Building on these insights, we introduced RoPE Profiler to measure both vulnerabilities from cached activations and guide targeted frequency interventions. Across the evaluated Qwen and Llama settings, we observed broadly consistent patterns of failure susceptibility, and our exploratory intervention search identified accuracy gains on a majority of settings without retraining. Together, these findings highlight the value of adapting frequency contributions to the needs of each task. We hope this work informs the design of long-context models that preserve relevant token preferences while resolving the positional distinctions a task requires.

AI USE STATEMENT

We used AI tools to aid or polish writing, adjust manuscript length and layout, create figures, organize the appendices, and search for related literature. We reviewed all generated sentences and figures, and checked against original sources for the validity of related literature. We take responsibility for the final content of this work.

REFERENCES

Christoph Aistleitner, István Berkes, and Robert Tichy. Lacunary sequences in analysis, probability and number theory. In Dijana Kreso, Joël Rivat, and Robert F. Tichy (eds.), *Diophantine Problems: Determinism, Randomness and Applications*, volume 62 of *Panoramas et Synthèses*, pp. 1–60. Société Mathématique de France, Paris, France, 2024. URL https://smf.emath.fr/publications/problemes-diophantiens-determinisme-alea-et-applications.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-Eval: Instituting standardized evaluation for long context language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.776. URL https://aclanthology.org/2024.acl-long.776/.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. Why does the effective context length of LLMs fall short? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 54791–54810, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/884baf65392170763b27c914087bde01-Paper-Conference.pdf.

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. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Rachit Bansal, Aston Zhang, Rishabh Tiwari, Lovish Madaan, Venkata Sai Surya Subramanyam Duvvuri, Devvrit Khatri, David Brandfonbrener, David Alvarez-Melis, Prajjwal Bhargava, Mihir Kale, and Samy Jelassi. Let's (not) just put things in context: Test-time training for long-context LLMs. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 112130–112153, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b619cd6dcc986856b8a8da2b08d89396-Paper-Conference.pdf.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! What makes rotary positional encodings useful? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 92408–92438, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e6d58fc68c0f3c36ae6e0e64478a69c0-Paper-Conference.pdf.

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

Lihu Chen, Gaël Varoquaux, and Fabian M. Suchanek. The locality and symmetry of positional encodings. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14313–14331, Singapore, December 2023a.

Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.955. URL https://aclanthology.org/2023.findings-emnlp.955/.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 86454–86478, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8b08bbf8b420faa6eeb4020720582ec7-Paper-Conference.pdf.

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

Yiting Chen and Junchi Yan. What rotary position embedding can tell us: Identifying query and key weights corresponding to basic syntactic or high-level semantic information. In *Advances in Neural Information Processing Systems*, volume 37, pp. 54507–54528. Curran Associates, Inc., 2024. doi: 10.52202/079017-1727. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/61f425da6e0a201b8fe1454601abfba5-Paper-Conference.pdf.

Yuhan Chen, Ang Lv, Jian Luan, Bin Wang, and Wei Liu. HoPE: A novel positional encoding without long-term decay for enhanced context awareness and extrapolation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23044–23056, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1123. URL https://aclanthology.org/2025.acl-long.1123/.

Ting-Rui Chiang and Dani Yogatama. The rotary position embedding may cause dimension inefficiency in attention heads for long-distance retrieval. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13552–13562, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.697. URL https://aclanthology.org/2025.findings-acl.697/.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11091–11104. PMLR, 2024. URL https://proceedings.mlr.press/v235/ding24i.html.

Zican Dong, Junyi Li, Xin Men, Wayne Xin Zhao, Bingning Wang, Zhen Tian, Weipeng Chen, and Ji-Rong Wen. Exploring context window of large language models via decomposed positional vectors. 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. 10320–10347. Curran Associates, Inc., 2024. doi: 10.52202/079017-0330. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1403ab1a427050538ec59c7f570aec8b-Paper-Conference.pdf.

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.

Yufeng Du, Phillip Harris, Minyang Tian, Eliu A Huerta, Srikanth Ronanki, Subendhu Rongali, Aram Galstyan, and Hao Peng. RoPE distinguishes neither positions nor tokens in long contexts, provably. arXiv preprint arXiv:2605.15514, 2026. URL https://arxiv.org/abs/2605.15514.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783v1.

Gemma Team. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025. URL https://arxiv.org/abs/2503.19786.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. Is it really long context if all you need is retrieval? towards genuinely difficult long context NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16576–16586, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Zihan Gu, Ruoyu Chen, Han Zhang, Hua Zhang, and Yue Hu. Deconstructing positional information: From attention logits to training biases. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=D0u0glT060.

Chi Han and Heng Ji. Computation mechanism behind LLM position generalization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19408–19424, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.953. URL https://aclanthology.org/2025.acl-long.953/.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.99. URL https://aclanthology.org/2022.findings-emnlp.99/.

Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On the token distance modeling ability of higher RoPE attention dimension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5877–5888, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.338. URL https://aclanthology.org/2024.findings-emnlp.338/.

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? In *Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=kIoBbc76Sy.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.890. URL https://aclanthology.org/2024.findings-acl.890/.

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24932–24949. PMLR, 2025.

Selim Jerad, Anej Svete, Jiaoda Li, and Ryan Cotterell. Disentangling the expressivity of RoPE. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2608.11909.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. Massive values in self-attention modules are the key to contextual knowledge understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28063–28096. PMLR, 2025.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, 2023.

Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=09-528y2Fgf.

Kimi Team. Kimi K3: Open frontier intelligence. arXiv preprint arXiv:2607.24653, 2026. URL https://arxiv.org/abs/2607.24653.

Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2676–2686, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1249. URL https://aclanthology.org/P18-1249/.

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. 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. 106519–106554. Curran Associates, Inc., 2024. doi: 10.52202/079017-3381. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c0d62e70dbc659cc9bd44cbcf1cb652f-Paper-Datasets_and_Benchmarks_Track.pdf.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.818. URL https://aclanthology.org/2024.acl-long.818/.

Haoran Li, Sucheng Ren, Alan Yuille, and Feng Wang. CoPE: Clipped RoPE as a scalable free lunch for long context LLMs. arXiv preprint arXiv:2602.05258, 2026. URL https://arxiv.org/abs/2602.05258.

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. LongReason: A synthetic long-context reasoning benchmark via context expansion. arXiv preprint arXiv:2501.15089, 2025. URL https://arxiv.org/abs/2501.15089.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3. arXiv preprint arXiv:2601.08584, 2026. URL https://arxiv.org/abs/2601.08584.

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

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, 2024a. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of RoPE-based extrapolation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 48160–48185, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d33b741600f100b31256c70a46f66ec9-Paper-Conference.pdf.

Yi Lu, Wanxu Zhao, Xin Zhou, Chenxin An, Chenglong Wang, Shuo Li, Yuming Yang, Jun Zhao, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Effective length extrapolation via dimension-wise positional embeddings manipulation. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Tahpc3iAnO.

Aryan Mikaeili, Or Patashnik, Andrea Tagliasacchi, Daniel Cohen-Or, and Ali Mahdavi-Amiri. Untwisting RoPE: Frequency control for shared attention in DiTs. arXiv preprint arXiv:2602.05013, 2026. URL https://arxiv.org/abs/2602.05013.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44554–44570. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/modarressi25a.html.

H. L. Montgomery and R. C. Vaughan. Hilbert's inequality. *Journal of the London Mathematical Society*, s2-8(1):73–82, May 1974. ISSN 0024-6107. doi: 10.1112/jlms/s2-8.1.73. URL https://doi.org/10.1112/jlms/s2-8.1.73.

Yui Oka, Kentaro Hanafusa, Taku Hasegawa, Kyosuke Nishida, and Kuniko Saito. Probing rotary position embeddings through frequency entropy. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 6139–6167, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0aee38a6fe9fffc8b658cfb1d872c1d5-Paper-Conference.pdf.

Yui Oka, Itsumi Saito, Kyosuke Nishida, and Kuniko Saito. Frequency bands in RoPE: Base frequency and context length shape the interpolation–extrapolation trade-off. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 94947–94975, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/993dbaec0418a6449090f1debbcb8844-Paper-Conference.pdf.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/874a4d89f2d04b4bcf9a2c19545cf040-Paper-Conference.pdf.

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*, 2022a. URL https://openreview.net/forum?id=R8sQPpGCv0.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022b.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2503.17349.

Ye Qiao, Haocheng Xu, Xiaofan Zhang, and Sitao Huang. Rethinking RoPE scaling in quantized LLM: Theory, outlier, and channel-band analysis with weight rescaling. arXiv preprint arXiv:2510.00028, 2025. URL https://arxiv.org/abs/2510.00028.

R. Salem and A. Zygmund. On lacunary trigonometric series. *Proceedings of the National Academy of Sciences of the United States of America*, 33(11):333–338, 1947. doi: 10.1073/pnas.33.11.333. URL https://doi.org/10.1073/pnas.33.11.333.

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

Runlin Shi, Bojian Yin, and Guoqi Li. Modern transformers are implicit hybrids: From functional differentiation to principled hybrid architecture design. arXiv preprint arXiv:2609.02986, 2026. URL https://arxiv.org/abs/2609.02986.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Bowen Sun, Yujun Cai, Ming-Hsuan Yang, Hang Wu, and Yiwei Wang. PAS: A training-free stabilizer for temporal encoding in video LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14480, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Sun_PAS_A_Training-Free_Stabilizer_for_Temporal_Encoding_in_Video_LLMs_CVPR_2026_paper.html.

Rongyuan Tan, Jue Zhang, Zhuozhao Li, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Contrastive attribution in the wild: An interpretability analysis of LLM failures on realistic benchmarks. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2026. URL https://arxiv.org/abs/2604.17761. To appear.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Danning Ke, Shikuan Hong, Yiwu Yao, and Gongyi Wang. RazorAttention: Efficient KV cache compression through retrieval heads. In *International Conference on Learning Representations*, 2025.

Felipe Urrutia, Jorge Salas, Alexander Kozachinskiy, Cristian Buc Calderon, Hector Pasten, and Cristobal Rojas. Decoupling positional and symbolic attention behavior in transformers. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 64123–64153, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/68f1a6933797528dd1ec8d15c6f58f62-Paper-Conference.pdf.

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*, pp. 5998–6008, 2017.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, Orhan Firat, Angeliki Lazaridou, Jean-Baptiste Lespiau, Nithya Attaluri, and Kate Olszewska. Michelangelo: Long context evaluations beyond haystacks via latent structure queries, 2024.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. Ada-LEval: Evaluating long-context LLMs with length-adaptable benchmarks. 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. 3712–3724, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.205. URL https://aclanthology.org/2024.naacl-long.205/.

Shaowen Wang, Yuke Zheng, Tansheng Zhu, Shuang Chen, Shaofan Liu, Suncong Zheng, and Jian Li. AdaRoPE: Not all attention heads should rotate and scale equally. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=xNuMtLaxo5.

Suyuchen Wang, Ivan Kobyzev, Peng Lu, Mehdi Rezagholizadeh, and Bang Liu. Resonance RoPE: Improving context length generalization of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 586–598, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.32. URL https://aclanthology.org/2024.findings-acl.32/.

Davis Wertheimer, Aozhong Zhang, Derrick Liu, Penghang Yin, and Naigang Wang. Frayed RoPE and long inputs: A geometric perspective. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 156004–156035, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fc684a13a24ac4735b1726659ce01049-Paper-Conference.pdf.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 62143–62156, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9b77f07301b1ef1fe810aae96c12cb7b-Paper-Conference.pdf.

Xiaodong Wu, Minhao Wang, Yichen Liu, Xiaoming Shi, He Yan, Xiangju Lu, Junmin Zhu, and Wei Zhang. LIFBench: Evaluating the instruction following performance and stability of large language models in long-context scenarios. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16445–16468, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.803. URL https://aclanthology.org/2025.acl-long.803/.

Xinyi Wu, Siyuan Liu, and Ali Jadbabaie. How data shapes RoPE frequency usage: From positional scale matching to length generalization. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026. URL https://arxiv.org/abs/2607.07678. Non-archival workshop paper.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5e5fd18f863cbe6d8ae392a93fd271c9-Paper-Conference.pdf.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. In *International Conference on Learning Representations*, 2025.

Jing Xiong, Liyang Fan, Hui Shen, Zunhai Su, Min Yang, Lingpeng Kong, and Ngai Wong. DoPE: Denoising rotary position embedding. arXiv preprint arXiv:2511.09146, 2025. URL https://arxiv.org/abs/2511.09146.

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 *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, 2024. Association for Computational Linguistics.

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, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024. URL https://arxiv.org/abs/2412.15115v1.

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

Bowen Yang, Bharat Venkitesh, Dwaraknath Gnaneshwar Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr Locatelli. Rope to Nope and back again: A new hybrid attention strategy. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 64133–64157. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2150. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5c9ab393551b7a39b4c02d88fe5e7e69-Paper-Conference.pdf.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context language models effectively and thoroughly. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 98914–98965, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f5332c8273d02729730a9c24dec2135e-Paper-Conference.pdf.

Justin Young. Effective harnesses for long-running agents. Anthropic Engineering, November 2025. URL https://www.anthropic.com/engineering/effective-harnesses-for-long-running-agents.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. 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. 60755–60775. Curran Associates, Inc., 2024. doi: 10.52202/079017-1943. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6ffdbbe354893979367f93e2121e37dd-Paper-Conference.pdf.

Zhisong Zhang, Yan Wang, Xinting Huang, Tianqing Fang, Hongming Zhang, Chenlong Deng, Shuaiyi Li, and Dong Yu. Attention entropy is a key factor: An analysis of parallel context encoding with full-attention-based pre-trained language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9840–9855, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.485. URL https://aclanthology.org/2025.acl-long.485/.

Mingming Zhao, Jiqian Dong, Kangping Xu, Zadid Hasan, Chengrui Fan, Shan Jiang, Shuai Mao, Yating Ling, Linyi Zou, Tailin Zhou, Yun Hin Chan, Wenkai Zhang, Zhanhong Zhou, Guowei Huang, Hongliang Li, Wenjing Cun, Zhitang Chen, Mingxuan Yuan, and Yanhui Geng. ScienceFlow: A long-horizon agent for ML research, scientific discovery and beyond. arXiv preprint arXiv:2608.14354, 2026. URL https://arxiv.org/abs/2608.14354.

Xinyu Zhao, Fangcong Yin, and Greg Durrett. Understanding synthetic context extension via retrieval heads. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77885–77910. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zhao25ad.html.

Shiyi Zhu, Jing Ye, Wei Jiang, Siqiao Xue, Qi Zhang, Yifan Wu, and Jianguo Li. CoCA: Fusing position embedding with collinear constrained attention in transformers for long context window extending. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4247–4262, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.233. URL https://aclanthology.org/2024.acl-long.233/.

Appendix

CONTENTS

A Formal Theory for the Main Results **21**

A.1 Notation for the Deferred Analysis 21

A.2 RoPE Score Reduction 22

A.3 Finite-Window High-Frequency Calibration 22

A.4 Positional Requirements 23

A.5 Precise Context-Length Theorem and Its Explicit Bound 23

B Extensions to Other Failure Modes **26**

B.1 Near–Far Ordering 26

B.2 Score-Barrier Stability 27

B.3 Fully-High Pairwise Reversal 29

B.4 Interpretation for p-RoPE 29

C Proofs of the Theoretical Results **30**

C.1 Proof of Proposition 2. 30

C.2 Finite-Window Fourier-Frame Bound 31

C.3 Proof of Proposition 1 and Lemma A.1 31

C.4 Strengthened Bandwise Calibration 33

C.5 Direct Finite-Sum Intuition 34

C.6 Positional Response Bounds and Refinements 34

C.7 Proof of Theorem 2 38

C.8 Proof of Lemma B.1. 39

C.9 Proof of Corollary B.1 40

C.10 Proof of Theorem 1 40

C.11 Proof of Theorem 3 42

D Empirical Validation of the Theory **43**

D.1 Local Positional-Response Envelope and Local-Response Floor 43

D.2 Experimental Details for Figure 3 44

D.3 Additional Evidence on High-Frequency Scores 45

E The 49 Reference Task Settings **47**

F Token and Head Selection **49**

F.1 Token Sampling from the Evaluated Inputs 49

F.2 Selecting a Fixed Set of Attention Heads 49

G Experimental Details **51**

G.1 Experiment 1: Task and Cross-Model Diagnostic Profiles 51

G.2 Experiment 2: Directed High-Frequency Intervention 53

H Results in the Assigned Optimization Direction **55**

H.1 Qwen3-8B: Semantic Direction 55

H.2 Qwen3-8B: Positional Direction 56

H.3 Llama-3.1-8B-Instruct: Semantic Direction 57

H.4 Llama-3.1-8B-Instruct: Positional Direction 58

I Additional Related Work **60**

I.1 Representation Assumptions in RoPE Theory 60

I.2 RoPE Failure Mechanisms. 60

I.3 Context Extension 60

I.4 Internal Representations and Failure Diagnosis 61

I.5 Behavioral Evaluation and Attention Competition 61

I.6 Long-Context Failure Modes 61

A Formal Theory for the Main Results

This appendix states the score representation, the finite-window moment guarantee, the local-response requirement, and the precise calibrated context-length theorem. Their proofs are collected in Appendix C; complementary failure modes appear in Appendix B.

A.1 Notation for the Deferred Analysis

We record the general notation used by the proofs and relate it to the high-frequency score S_H in §3.1. Consider a fully rotary head with $h \geq 1$ rotary coordinate pairs and base $B > 1$. Define

$$\rho := B^{-1/h}, \qquad \omega_n := \rho^n, \qquad n = 0, \ldots, h-1, \qquad \mathcal{F} := \{0, \ldots, h-1\}. \tag{8}$$

For fixed coefficients $z \in \mathbb{C}^h$ and a frequency band $J \subseteq \mathcal{F}$, write

$$X_J^z(m) := \mathrm{Re} \sum_{n \in J} z_n e^{im\omega_n}, \qquad J \subseteq \mathcal{F}. \tag{9}$$

Thus the fully rotary score $S(m)$ in the main text is $X_{\mathcal{F}}^z(m)$, and its high-frequency contribution is $S_H(m) = X_H^z(m)$. The general band norm is

$$U_J(z) := \left(\sum_{n \in J} |z_n|^2 \right)^{1/2}, \qquad U(z) := U_{\mathcal{F}}(z) = \|z\|_2. \tag{10}$$

For an integer window length $M \geq 2$, let

$$\langle f \rangle_M := \frac{1}{M} \sum_{m=0}^{M-1} f(m). \tag{11}$$

The notation $\mathrm{Var}_M(f)$ denotes $\left\langle (f - \langle f \rangle_M)^2 \right\rangle_M$ for real-valued f. Equivalently, these are the mean and variance under $\mathbf{m} \sim \mathrm{Unif}\{0, \ldots, M-1\}$.

Let $C_{\mathrm{F}} > 0$ be the absolute Fourier-frame constant in Lemma C.1. For a fixed tolerance $0 < \varepsilon < 1/2$, define

$$\Gamma_\varepsilon(\rho) := \frac{2C_{\mathrm{F}}}{(1-\rho)\varepsilon}, \qquad \mathcal{H}_\varepsilon(M) := \{n \in \mathcal{F} : M\omega_n \geq \Gamma_\varepsilon(\rho)\}, \qquad \mathcal{N}_\varepsilon(M) := \mathcal{F} \backslash \mathcal{H}_\varepsilon(M). \tag{12}$$

A certified window satisfies $M \geq \Gamma_\varepsilon(\rho)$, so the high-frequency prefix $H = \mathcal{H}_\varepsilon(M)$ is nonempty. Its complement is $N = \mathcal{N}_\varepsilon(M)$. The random variable used in the calibration proof is

$$Y_H^z := X_H^z(\mathbf{m}) = S_H(\mathbf{m}), \qquad \mathbf{m} \sim \mathrm{Unif}\{0, \ldots, M-1\}. \tag{13}$$

For $z \neq 0$, the two norm shares are

$$r_H(z; M) := \frac{U_H(z)}{U(z)}, \qquad r_N(z; M) := \frac{U_N(z)}{U(z)}, \qquad r_H(z; M)^2 + r_N(z; M)^2 = 1. \tag{14}$$

These coincide with the main-text high-frequency norm share and its complement. Their squares are the corresponding energy fractions.

Apply the normalized adjacent response from Eq. (3) to $X_{\mathcal{F}}^z$. Failure to attain response level $\zeta > 0$ anywhere in the window means

$$\max_{0 \leq m \leq M-2} g_z(m) = \max_{0 \leq m \leq M-2} \frac{|X_{\mathcal{F}}^z(m+1) - X_{\mathcal{F}}^z(m)|}{U(z)} < \zeta. \tag{15}$$

For the ordered margin $D(m) = X_{\mathcal{F}}^d(m)$ with reference ordering $D(0) > 0$, the coefficient-based and function-based reversal notation refer to the same probability:

$$p_{\mathrm{rev}}(d; M) = p_{\mathrm{rev}}(D; M) := \mathbb{P}[D(\mathbf{m}) < 0] = \mathbb{P}[X_{\mathcal{F}}^d(\mathbf{m}) < 0]. \tag{16}$$

In particular, the main theorem's g_d is the normalized adjacent response of this same margin, with denominator $U(d)$.

A.2 RoPE Score Reduction

The reduction uses RoPE's relative-rotation identity (Su et al., 2024).

Proposition 2 (RoPE score reduction)**.** *For a fully rotary RoPE head, as the real query and key blocks range over $\mathbb{R}^{2h}$, the class of fixed query–key scores and their finite real linear combinations as functions of relative distance is exactly $\{X^z_{\mathcal{F}} : z \in \mathbb{C}^h\}$; hence all subsequent full- and bandwise analyses reduce to studying X^z_j.*

Complex-coordinate derivation. Write the two real coordinates of each rotary pair as $Q_n = q_{n,1} + iq_{n,2}$ and $K_n = k_{n,1} + ik_{n,2}$, and let $R(\theta)$ denote their planar rotation. For relative distance $m = p_q - p_k$ and a position-independent score scale $c_{\mathrm{att}} > 0$,

$$\begin{aligned} c_{\mathrm{att}} \sum_n \langle R(p_q\omega_n)q_n, R(p_k\omega_n)k_n\rangle &= c_{\mathrm{att}} \sum_n q_n^\top R(-m\omega_n)k_n \\ &= \mathrm{Re} \sum_n \underbrace{c_{\mathrm{att}} Q_n \overline{K_n}}_{z_n} e^{im\omega_n} = X^z_{\mathcal{F}}(m). \end{aligned}$$

The same calculation applies to score margins by subtracting their coefficient vectors. Appendix C.1 gives the full derivation and the converse construction.

Operationally, the reduction isolates all dependence on relative distance in a finite trigonometric polynomial. It preserves the spectral competition between the diagnostics while leaving content-state changes and downstream attention computations outside the certificate.

Under the same fixed-content assumption, the argument does not depend on whether an implementation stores each rotary pair contiguously or in a split-half layout; only the pairing of real coordinates matters. A non-rotary tail then contributes a position-independent constant. It is omitted from the formal fully rotary results, while Appendix B.4 discusses its distinct effects on cross-key score margins and same-key positional comparisons.

A.3 Finite-Window High-Frequency Calibration

The moment conclusion below is Proposition 1 in the notation of this appendix. The additional conditional statement records the Gaussian approximation error used by the extensions.

Lemma A.1 (Finite-window moments and conditional Gaussian calibration)**.** *Fix $0 < \varepsilon < 1/2$, a certified window $M \geq \Gamma_\varepsilon(\rho)$, and fixed coefficients z with $U_H(z) > 0$. Use $H = \mathcal{H}_\varepsilon(M)$ from* Eq. (12) *and the uniformly sampled relative distance in* Eq. (13)*. For its random score Y^z_H, write*

$$\mu_H := \langle X^z_H \rangle_M, \qquad \sigma^2_H := \mathrm{Var}_M(X^z_H). \tag{17}$$

Then the following moment bounds hold:

$$|\mu_H| \leq \varepsilon U_H(z), \qquad \left|\sigma^2_H - \tfrac{1}{2}U_H(z)^2\right| \leq \varepsilon U_H(z)^2. \tag{18}$$

Because $0 < \varepsilon < 1/2$, Eq. (18) *implies $\sigma^2_H \geq (1/2 - \varepsilon)U_H(z)^2 > 0$ whenever $U_H(z) > 0$. The standardization below is therefore well defined. The remaining, non-deterministic source of approximation error is the Gaussian-shape discrepancy of the standardized window law,*

$$\Delta_{\mathrm{H,G}}(z; M) := \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left[\frac{Y^z_H - \mu_H}{\sigma_H} \leq x\right] - \Phi_{\mathrm{std}}(x)\right| \leq \tau_{\mathrm{H,G}}. \tag{19}$$

Under this condition, the exact form of the Gaussian conclusion is

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}[Y^z_H \leq x] - \Phi_{\mathrm{std}}\left(\frac{\sqrt{2}\,x}{U_H(z)}\right)\right| \leq \tau_{\mathrm{H,G}} + c_{\mathrm{G}}(\varepsilon), \tag{20}$$

where the universal moment-replacement term is

$$c_{\mathrm{G}}(\varepsilon) := \frac{\varepsilon}{\sqrt{\pi}} + \frac{\log((1-2\varepsilon)^{-1})}{2\sqrt{2\pi e}} = O(\varepsilon) \qquad (\varepsilon \to 0). \tag{21}$$

We use the convention $\mathcal{N}(\mu, \sigma^2)$: the target variance is $U_H(z)^2/2$, equivalently the target standard deviation is $U_H(z)/\sqrt{2}$.

A.4 Positional Requirements

A.4.1 Local Positional Response

For a head whose function requires neighboring relative positions to remain distinguishable, moving a key by one token should induce a non-negligible normalized score change. In the terminology of §3.2, avoiding the window-level failure in Eq. (15) at response level ζ requires at least one adjacent score change of normalized magnitude ζ. We now quantify how maintaining this response constrains the certified-high norm share as the context scale grows.

Lemma A.2 (Direct scale-to-share response envelope)**.** *For every certified integer $M \geq 2$ and every $a \neq 0$,*

$$\max_{0\leq m\leq M-2} \frac{|X^a_{\mathcal{F}}(m+1) - X^a_{\mathcal{F}}(m)|}{U(a)} \leq \underbrace{\frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{M}\, r_N(a;M)}_{\substack{\text{non-high band:}\\ \text{small rotations}}} + \underbrace{R_{\mathcal{F}}\, r_H(a;M)}_{\substack{\text{certified-high band:}\\ \text{coefficient norm}}} \leq \frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{M} + R_{\mathcal{F}}\, r_H(a;M). \tag{22}$$

where $R_{\mathcal{F}} := \left(\sum_{n\in\mathcal{F}} \left|e^{i\omega_n} - 1\right|^2\right)^{1/2} > 0$ is the fixed full-grid adjacent-gain factor and is independent of M.

The non-high contribution is bounded by a $1/M$ envelope because every non-high frequency rotates by less than $\Gamma_\varepsilon(\rho)/M$ per token. The second term has no explicit inverse-scale decay: its full-grid gain factor is independent of M, but the certified-high norm share can still vary with the window.

Corollary A.1 (Stable local-response floor on the certified-high norm share)**.** *For every certified integer $M \geq 2$, every $a \neq 0$, and every $\zeta > 0$, if the score $X^a_{\mathcal{F}}$ avoids the window-level failure in* Eq. (15) *at response level ζ, then*

$$r_H(a;M) \geq \underline{r}_{\text{pos}}(M;\zeta) := \frac{\left[\zeta - \Gamma_\varepsilon(\rho)\sqrt{h}/M\right]_+}{R_{\mathcal{F}}}. \tag{23}$$

where $R_{\mathcal{F}}$ is the fixed factor in Lemma A.2. If the right-hand side exceeds one, the requested response is impossible.

Together, Lemma A.2 and Corollary A.1 directly link the context scale M to the certified-high norm share: maintaining the same adjacent response level ζ forces $r_H(a;M)$ above $\underline{r}_{\text{pos}}(M;\zeta)$. This floor is the restriction to certified integer windows of a continuous, non-decreasing function of real $M > 0$, and it is strictly increasing once positive. Thus it provides a stable non-decreasing lower bound even though $H(M)$ changes only at discrete breakpoints. Enlarging $H(M)$ alone would not give this conclusion, since a newly certified coefficient may be zero and the actual share need not jump. The continuity and monotonicity claim concerns the necessary floor, not the actual stepwise quantity $r_H(a;M)$ or its increment between two windows. No prior assumption $U_H(a) > 0$ is needed: a positive floor itself forces that conclusion. Sharper coefficient-specific refinements are given in Appendix C.6.4.

The response floor applies to the same ordered margin used in the context-length result below.

A.5 Precise Context-Length Theorem and Its Explicit Bound

Fix a fully rotary head, $0 < \varepsilon < 1/2$, and the same nonzero ordered margin coefficients d with $D(0) > 0$ across a finite ordered set of certified windows $\mathfrak{M}_{\text{cal}}$. Fix tolerances $0 \leq \delta < 1/2$ and $0 < \zeta < R_{\mathcal{F}}$, together with a uniform Gaussian threshold-error bound $0 \leq \overline{\tau}_{\text{G}} < 1/2$. For the ordered margin d in §3.2, set

$$\mu_D(M) := \left\langle X^d_{\mathcal{F}} \right\rangle_M, \qquad \sigma^2_D(M) := \text{Var}_M(X^d_{\mathcal{F}}), \tag{24}$$

and assume $\sigma_D(M) > 0$ at every window in $\mathfrak{M}_{\text{cal}}$. The exact-moment Gaussian threshold error is

$$\Delta_{\text{rev}}(d;M) := \left| p_{\text{rev}}(d;M) - \Phi_{\text{std}}\left(-\frac{\mu_D(M)}{\sigma_D(M)}\right)\right|. \tag{25}$$

The calibrated family in Theorem 1 satisfies $\Delta_{\mathrm{rev}}(d; M) \le \overline{\tau}_{\mathrm{G}}$ uniformly.

Let

$$k_N(M) := |\mathcal{N}_\varepsilon(M)|\,, \qquad c_\varepsilon := \sqrt{\frac{1}{2} - \varepsilon}, \tag{26}$$

and, for $0 \le r \le 1$, define

$$\begin{aligned} u_M(r) &:= \varepsilon r + \sqrt{k_N(M)}\sqrt{1 - r^2}, \\ v_M(r) &:= \left[c_\varepsilon r - \sqrt{k_N(M)}\sqrt{1 - r^2}\right]_+ . \end{aligned} \tag{27}$$

Recall the full-grid adjacent-gain bound $R_{\mathcal{F}} = (\sum_{n\in\mathcal{F}} |e^{i\omega_n} - 1|^2)^{1/2} > 0$. The main theorem assumes $0 < \zeta < R_{\mathcal{F}}$. By Corollary A.1, retaining response level ζ for the same margin d requires

$$r_H(d; M) \ge \ell(M; \zeta) := \frac{[\zeta - \Gamma_\varepsilon(\rho)\sqrt{h}/M]_+}{R_{\mathcal{F}}}. \tag{28}$$

Here $[x]_+ = \max\{x, 0\}$. The stated range of ζ ensures $0 \le \ell(M; \zeta) < 1$, within the domain of the reversal envelope.

The Gaussian reversal envelope and the floor used in the main text are

$$\psi_M(r; \overline{\tau}_{\mathrm{G}}) := \begin{cases} [\Phi_{\mathrm{std}}(-u_M(r)/v_M(r)) - \overline{\tau}_{\mathrm{G}}]_+\,, & v_M(r) > 0, \\ 0, & v_M(r) = 0, \end{cases} \tag{29}$$

$$\underline{p}_{\mathrm{rev}}(M; \zeta, \overline{\tau}_{\mathrm{G}}) := \psi_M(\ell(M; \zeta); \overline{\tau}_{\mathrm{G}})\,. \tag{30}$$

For fixed ζ and $\overline{\tau}_{\mathrm{G}}$, we abbreviate this bound as $\underline{p}_{\mathrm{rev}}(M)$.

For every calibrated window at which the margin retains response level ζ, the lower-bound chain underlying Theorem 1 is

$$p_{\mathrm{rev}}(d; M) \ge \psi_M(r_H(d; M); \overline{\tau}_{\mathrm{G}}) \ge \psi_M(\ell(M; \zeta); \overline{\tau}_{\mathrm{G}}) = \underline{p}_{\mathrm{rev}}(M; \zeta, \overline{\tau}_{\mathrm{G}}). \tag{31}$$

Appendix C.10 proves this chain and its monotonicity in M.

When the crossing set is nonempty, the first excluded scale is

$$M_\dagger := \min\left\{M \in \mathfrak{M}_{\mathrm{cal}} : \underline{p}_{\mathrm{rev}}(M; \zeta, \overline{\tau}_{\mathrm{G}}) > \delta\right\}. \tag{32}$$

For every $M \in \mathfrak{M}_{\mathrm{cal}}$ with $M \ge M_\dagger$,

$$p_{\mathrm{rev}}(D; M) > \delta \qquad \text{or} \qquad \max_{0\le m\le M-2} g_d(m) < \zeta. \tag{33}$$

If the crossing set is empty, the bound supplies no exclusion threshold within the chosen calibrated family.

If the sub-threshold set is nonempty, define

$$M_{\max}^{\mathrm{cert}} := \max\left\{M \in \mathfrak{M}_{\mathrm{cal}} : \underline{p}_{\mathrm{rev}}(M; \zeta, \overline{\tau}_{\mathrm{G}}) \le \delta\right\}. \tag{34}$$

This is the largest calibrated scale not ruled out by the lower bound. When the crossing set in Eq. (32) is also nonempty, it is the grid point immediately preceding $M_\dagger$. The main text writes $M_{\max} = M_{\max}^{\mathrm{cert}}$ and uses the regime in which both the crossing and sub-threshold sets are nonempty. For calibrated windows, $M > M_{\max}$ is then equivalent to $M \ge M_\dagger$. All such windows fail at least one requirement, which gives the simultaneous-infeasibility statement in Eq. (7). A scale below the crossing remains a candidate: the lower bound alone does not establish joint reliability there. This count-based envelope is deliberately conservative. It uses no coefficient information beyond the certified-high share; the practical predictor in Appendix G.1.2 computes the exact full-window mean and variance, including cross-band covariance.

Theorem 1 (Precise context-length limit on joint reliability)**.** *Fix a fully rotary head with the geometric frequency grid in* Eq. (8)*, a tolerance $0 < \varepsilon < 1/2$, and a nonzero ordered margin $D(m) = X^d_{\mathcal{F}}(m)$ with $D(0) > 0$. Let $\mathfrak{M}_{\mathrm{cal}}$ be a finite nonempty set of certified integer windows $M \geq 2$, with the same coefficients d at every window. Fix $0 \leq \delta < 1/2$, $0 < \zeta < R_{\mathcal{F}}$, and $0 \leq \overline{\tau}_{\mathrm{G}} < 1/2$. Assume, for every $M \in \mathfrak{M}_{\mathrm{cal}}$, that $\sigma_D(M) > 0$ and*

$$\Delta_{\mathrm{rev}}(d; M) = \left| p_{\mathrm{rev}}(d; M) - \Phi_{\mathrm{std}}\left(-\frac{\mu_D(M)}{\sigma_D(M)} \right) \right| \leq \overline{\tau}_{\mathrm{G}}.$$

Let $\underline{p}_{\mathrm{rev}}(M; \zeta, \overline{\tau}_{\mathrm{G}})$ be the explicit bound in Eq. (30)*. This bound is non-decreasing across $\mathfrak{M}_{\mathrm{cal}}$, and each calibrated window retaining the adjacent response $\max_{0 \leq m \leq M-2} g_d(m) \geq \zeta$ satisfies*

$$p_{\mathrm{rev}}(d; M) \geq \underline{p}_{\mathrm{rev}}(M; \zeta, \overline{\tau}_{\mathrm{G}}).$$

Write $M_- := \min \mathfrak{M}_{\mathrm{cal}}$ and $M_+ := \max \mathfrak{M}_{\mathrm{cal}}$. If the reversal tolerance satisfies

$$\underline{p}_{\mathrm{rev}}(M_-) \leq \delta < \underline{p}_{\mathrm{rev}}(M_+), \tag{35}$$

define $M_{\max} := M^{\mathrm{cert}}_{\max}$ by Eq. (34)*. Then, for every $M \in \mathfrak{M}_{\mathrm{cal}}$ with $M > M_{\max}$,*

$$p_{\mathrm{rev}}(d; M) > \delta \qquad \text{or} \qquad \max_{0 \leq m \leq M-2} g_d(m) < \zeta.$$

Thus $M_{\max}$ is an upper bound on joint score-level reliability within the calibrated family. A window at or below $M_{\max}$ remains a candidate; the bound alone does not establish joint reliability there.

An explicit sufficient condition for the strict upper inequality in Eq. (35) *is $M_+ \geq M_{\mathrm{suff}}$ and $0 \leq \delta < \delta_\star$, where*

$$\delta_\star := \left[\Phi_{\mathrm{std}}\left(-\frac{\varepsilon}{\sqrt{1/2 - \varepsilon}} \right) - \overline{\tau}_{\mathrm{G}} \right]_+, \tag{36}$$

$$\omega_{\min} := \rho^{h-1},$$

$$M_{\mathrm{suff}} := \max\left\{ 2, \left\lceil \frac{\Gamma_\varepsilon(\rho)}{\omega_{\min}} \right\rceil, \left\lfloor \frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{\zeta} \right\rfloor + 1 \right\}. \tag{37}$$

Indeed, $\underline{p}_{\mathrm{rev}}(M) \leq \delta_\star < 1/2$ at every calibrated window, with equality for each such window $M \geq M_{\mathrm{suff}}$. The lower inequality $\underline{p}_{\mathrm{rev}}(M_-) \leq \delta$ remains required.

Interpreting the reversal tolerance. The tolerance δ specifies the allowed fraction of relative distances at which the reference ordering reverses under uniform sampling. For example, $\delta = 0.38$ permits reversals at up to 38% of the positions in the window. Preserving reference orderings reliably calls for a much smaller tolerated fraction, so an upper restriction on δ is consistent with the intended semantic-stability requirement. The range in Eq. (35) identifies tolerances for which this bound certifies an exclusion threshold. Its use still requires the stated calibration conditions and the specified window family.

B Extensions to Other Failure Modes

The score representation in Appendix A also supports near–far comparisons of one key, exceedance of a fixed score barrier, fully-high pairwise reversal, and a partial-RoPE interpretation. Each result below retains its own event and calibration conditions. Proofs are collected in Appendix C.

B.1 Near–Far Ordering

B.1.1 Extended Positional Diagnostic: Near–Far Ordering

Adjacent response measures local resolution, but it does not show whether a head preserves a directional notion of distance at larger scales. For a head or task that requires monotone near preference, we orient the score so that a farther occurrence should receive a lower score. For an integer $M \geq 1$, independently sample m_{near} from $\{M+1, \ldots, 2M\}$ and m_{far} from the equally long, more distant interval $\{2M+1, \ldots, 3M\}$. For a fixed key K, define the ordering-error probability

$$p_{\mathrm{ord}}(K; M) := \mathbb{P}[S_K(m_{\mathrm{far}}) > S_K(m_{\mathrm{near}})]. \tag{38}$$

A reliable near–far order requires this probability to be small. When it is close to one half, the intended inequality is violated in half of the random comparisons under this audit protocol. We call this *near–far order ambiguity*; it is a score-level diagnostic of chance-like directional ordering in the two-interval comparison, not a claim that all positional information has disappeared.

This is a one-sided diagnostic: a small p_{ord} is necessary but not sufficient for reliable near preference, because a position-independent score also makes the strict event impossible. Fixed-offset, periodic, diagonal, and other non-monotone patterns may likewise remain detectable.

B.1.2 Near–Far Order Ambiguity

We apply the two-interval experiment from Appendix B.1.1 to one fixed key. The theorem below concerns the asymptotic fully-high calibration regime: when all rotating frequencies enter the certified-high band and the interval calibration errors vanish, the near–far ordering error approaches chance. Appendix B.1.3 gives the precise calibration conditions.

Theorem 2 (Near–far ordering approaches chance under full calibration)**.** *Fix a fully rotary head together with fixed query and key content states that induce a nonzero coefficient vector a. Independently draw m_{near} uniformly from $\{M+1, \ldots, 2M\}$ and m_{far} uniformly from $\{2M+1, \ldots, 3M\}$. Suppose the two interval score laws lie in the asymptotic fully-high calibration regime of Appendix B.1.3. Then*

$$p_{\mathrm{ord}}(a; M) := \mathbb{P}[X^a_{\mathcal{F}}(m_{\mathrm{far}}) > X^a_{\mathcal{F}}(m_{\mathrm{near}})] \longrightarrow \frac{1}{2} \qquad (M \to \infty). \tag{39}$$

The slowly rotating part can initially support a signed mean advantage for the nearer interval. As those frequencies enter the certified-high band, Proposition 1 drives both interval means toward zero and both variances toward the same half-energy value. Under the stated shape calibration, the two scores therefore approach the same distribution, so the farther score exceeds the nearer one with probability one half. The vanishing shape error remains substantive: equalized moments alone do not make a fixed finite trigonometric sum Gaussian.

B.1.3 Two-Interval Calibration Error

For an integer shift s, define

$$a_n^{[s]} := a_n e^{is\omega_n}, \qquad X^a_{\mathcal{F}}(s+t) = X^{a^{[s]}}_{\mathcal{F}}(t), \qquad U(a^{[s]}) = U(a). \tag{40}$$

The near and far intervals are length-M windows generated by $a^{[M+1]}$ and $a^{[2M+1]}$, respectively. Define

$$\tau_{\mathrm{near}}(a; M) := \Delta_{\mathrm{H,G}}(a^{[M+1]}; M), \qquad \tau_{\mathrm{far}}(a; M) := \Delta_{\mathrm{H,G}}(a^{[2M+1]}; M), \tag{41}$$

and

$$\tau_{\mathrm{ord}}(a; M, \varepsilon) := \tau_{\mathrm{near}}(a; M) + \tau_{\mathrm{far}}(a; M) + 2c_{\mathrm{G}}(\varepsilon). \tag{42}$$

For fixed finite h, set

$$\varepsilon_M := \frac{2C_{\mathrm{F}}}{(1-\rho)M\omega_{h-1}} \tag{43}$$

for all sufficiently large M. The asymptotic fully-high calibration regime in Theorem 2 imposes two requirements. Both standardized translated laws must satisfy Eq. (19) with $\varepsilon = \varepsilon_M$, and $\tau_{\mathrm{ord}}(a; M, \varepsilon_M) \to 0$. When available, valid held-out upper bounds on the two shape errors in Eq. (41) may be used in Eq. (42). This tolerance choice makes the partition fully high because $\Gamma_{\varepsilon_M}(\rho) = M\omega_{h-1}$, so $\mathcal{H}_{\varepsilon_M}(M) = \mathcal{F}$.

B.2 Score-Barrier Stability

B.2.1 Single-Key Exceedance Against a Score Barrier

We use rare exceedance of a fixed barrier as a single-key score-level certificate. It acquires a semantic ranking interpretation only under the winner-barrier construction below and its additional tail conditions.

Fix the query, layer, head, content states, and one key K with nonzero RoPE coefficient vector b. For a certified window, draw $\mathbf{m} \sim \mathrm{Unif}\{0, \dots, M-1\}$ and, for a fixed score barrier $S \in \mathbb{R}$, define

$$Y_b := X^b_{\mathcal{F}}(\mathbf{m}), \qquad p_{\mathrm{exc}}(b; S, M) := \mathbb{P}[Y_b > S]. \tag{44}$$

Write $Y_N := X^b_N(\mathbf{m})$ and $Y_H := X^b_H(\mathbf{m})$. Guided by Proposition 1, retain the non-high moments and the cross-band covariance, but replace the certified-high mean and variance by 0 and $U_H(b)^2/2$. This gives the Gaussian exceedance predictor

$$\begin{aligned} \widehat{\sigma}_b^2 &:= \mathrm{Var}_M(Y_N) + \frac{1}{2}U_H(b)^2 + 2\,\mathrm{Cov}_M(Y_N, Y_H), \\ \widehat{p}_{\mathrm{exc}}(b; S, M) &:= \Phi_{\mathrm{std}}\left(\frac{\langle X^b_N \rangle_M - S}{\widehat{\sigma}_b}\right), \end{aligned} \tag{45}$$

where Cov_M is covariance under the same uniform window and $\widehat{\sigma}_b$ is the positive square root. Let $\tau_{\mathrm{exc}}(b; S, M)$ denote the total calibration error defined in Appendix B.2.2. It combines the Gaussian-shape discrepancy of the full score with the high-band moment replacement error; the result below is informative when this measurable quantity is small.

Lemma B.1 (Calibrated single-key score exceedance)**.** *For every certified M, nonzero b, and $S \in \mathbb{R}$ in the non-degenerate calibration regime of Appendix B.2.2,*

$$|p_{\mathrm{exc}}(b; S, M) - \widehat{p}_{\mathrm{exc}}(b; S, M)| \le \tau_{\mathrm{exc}}(b; S, M). \tag{46}$$

Let $u_{\mathrm{exc}}(b; S, M, \delta)$ denote the explicit, coefficient-specific ceiling given in Appendix B.2.3. Its formula is deferred because only the resulting bound, rather than the quadratic inversion used to obtain it, is needed here.

Corollary B.1 (Small score-barrier exceedance imposes a certified-high norm-share ceiling)**.** *Under the hypotheses of Lemma B.1, assume $U_N(b)U_H(b) > 0$ and $0 < \delta + \tau_{\mathrm{exc}}(b; S, M) < 1/2$. Then*

$$p_{\mathrm{exc}}(b; S, M) \le \delta \quad \Longrightarrow \quad r_H(b; M) \le u_{\mathrm{exc}}(b; S, M, \delta) < 1. \tag{47}$$

This is exactly the same $r_H(b; M) = U_H(b)/U(b)$ constrained from below by the local-response floor when it is applied to b. Hence, within the calibration and interior hypotheses above, a local-response floor above u_{exc} forces $p_{\mathrm{exc}}(b; S, M) > \delta$. The endpoint cases are different: $U_H(b) = 0$ gives $r_H = 0$ directly, whereas at $U_N(b) = 0$ rare exceedance of a sufficiently high barrier need not imply any ceiling below one.

Winner-barrier interpretation. Let $\mathcal{K}$ be a finite key bank at the reference placement $m = 0$, choose any reference winner, and define

$$K_+ \in \arg\max_{K_j \in \mathcal{K}} X^{b_j}_{\mathcal{F}}(0), \qquad S_{\max} := X^{b_+}_{\mathcal{F}}(0), \qquad \alpha_+(S_{\max}; M) := \mathbb{P}[X^{b_+}_{\mathcal{F}}(\mathbf{m}) > S_{\max}]. \tag{48}$$

For every competitor $K_j \neq K_+$, event inclusion gives

$$p_{\mathrm{rev}}(b_+, b_j; M) := \mathbb{P}[X_{\mathcal{F}}^{b_j}(\mathbf{m}) > X_{\mathcal{F}}^{b_+}(\mathbf{m})] \geq [p_{\mathrm{exc}}(b_j; S_{\max}, M) - \alpha_+(S_{\max}; M)]_+ . \tag{49}$$

Consequently, fix $q > 0$ and suppose $\alpha_+(S_{\max}; M) \leq \delta - q$. If every competitor lies in the calibrated interior regime above and its local-response floor exceeds $u_{\mathrm{exc}}(b_j; S_{\max}, M, \delta)$, then each competitor has the marginal guarantee $p_{\mathrm{rev}}(b_+, b_j; M) > q$. Thus choosing S as the largest reference score exposes a competitor-wise marginal reversal risk across the bank once all candidate-specific intervals are incompatible. It does not assert that all competitors reverse at the same relative distance, nor does the maximum-barrier choice alone imply the result: the calibration, interval crossing, and winner-tail condition must all hold.

The score-barrier ceiling above applies in the calibrated interior regime. The fully-high endpoint is covered by the ordered-pair result in Appendix B.3.

B.2.2 Calibration Quantities

For the fixed score in §B.2.1, write $Y_N := X_N^b(\mathbf{m})$, $Y_H := X_H^b(\mathbf{m})$, and set

$$\begin{aligned} \mu_b &:= \left\langle X_{\mathcal{F}}^b \right\rangle_M , & \sigma_b^2 &:= \mathrm{Var}_M(X_{\mathcal{F}}^b), \\ \mu_N &:= \left\langle X_N^b \right\rangle_M , & \sigma_N^2 &:= \mathrm{Var}_M(X_N^b), \\ \kappa_{NH} &:= \left\langle (X_N^b - \mu_N)(X_H^b - \left\langle X_H^b \right\rangle_M) \right\rangle_M , \\ \widehat{\sigma}_b^2 &:= \sigma_N^2 + \frac{1}{2} U_H(b)^2 + 2\kappa_{NH}. \end{aligned} \tag{50}$$

Thus $\widehat{\sigma}_b^2$ is exactly the moment-calibrated variance proxy in Eq. (45). The non-degenerate calibration regime used in Lemma B.1 is

$$\sigma_b^2 > 0, \qquad \widehat{\sigma}_b^2 > \varepsilon U_H(b)^2. \tag{51}$$

In this regime, define the full-score Gaussian-shape discrepancy

$$\Delta_{\mathrm{exc}}(b; M) := \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left[\frac{Y_b - \mu_b}{\sigma_b} \leq x \right] - \Phi_{\mathrm{std}}(x) \right| \tag{52}$$

and the variance floor and total calibration error

$$\begin{aligned} s_-(b; M) &:= \sqrt{\widehat{\sigma}_b^2 - \varepsilon U_H(b)^2}, \\ \tau_{\mathrm{exc}}(b; S, M) &:= \Delta_{\mathrm{exc}}(b; M) + \frac{1}{\sqrt{2\pi}} \left[\frac{\varepsilon U_H(b)}{s_-(b; M)} + \frac{|S - \mu_N| \, \varepsilon U_H(b)^2}{s_-(b; M) \widehat{\sigma}_b \big(s_-(b; M) + \widehat{\sigma}_b \big)} \right]. \end{aligned} \tag{53}$$

The first term measures the deviation of the *full* score law from Gaussianity; the remaining terms account for replacing the high-band moments using Proposition 1.

B.2.3 Explicit Ceiling Candidate

For the interior case $U_N(b)U_H(b) > 0$ and $0 < \delta + \tau_{\mathrm{exc}}(b; S, M) < 1/2$, define

$$\begin{aligned} \lambda_S &:= \frac{S - \mu_N}{U_N(b)}, & \beta_N &:= \frac{\sigma_N^2}{U_N(b)^2}, \\ \gamma_{NH} &:= \frac{\kappa_{NH}}{U_N(b)U_H(b)}, & \xi_{\mathrm{exc}} &:= \Phi_{\mathrm{std}}^{-1}(1 - \delta - \tau_{\mathrm{exc}}(b; S, M)) . \end{aligned} \tag{54}$$

The corresponding explicit ceiling candidate is

$$\begin{aligned} \bar{t}_{\mathrm{exc}} &:= -2\gamma_{NH} + \sqrt{4\gamma_{NH}^2 + 2\left(\frac{\lambda_S^2}{\xi_{\mathrm{exc}}^2} - \beta_N \right)}, \\ u_{\mathrm{exc}}(b; S, M, \delta) &:= \frac{\bar{t}_{\mathrm{exc}}}{\sqrt{1 + \bar{t}_{\mathrm{exc}}^2}}. \end{aligned} \tag{55}$$

Its radicand need not be nonnegative for arbitrary inputs. Under the antecedent of Corollary B.1, the induced high-to-non-high norm ratio is feasible for the resulting quadratic; this guarantees that the candidate is real and positive.

B.3 Fully-High Pairwise Reversal

For an ordered key pair, we call the event that the first score falls below the second a directional reversal; it is semantic when the first key is the preferred candidate.

Theorem 3 (Fully-high pairwise reversal approaches chance)**.** *For a fully rotary head, fix a query and an ordered pair of keys with coefficient vectors $a, b \in \mathbb{C}^h$ such that $d := a - b \neq 0$. For any $0 < \varepsilon < 1/2$ and any certified window satisfying $\mathcal{H}_\varepsilon(M) = \mathcal{F}$, draw $\mathbf{m} \sim \mathrm{Unif}\{0, \dots, M-1\}$. If the Gaussian-shape condition in* Eq. (19) *holds for $z = d = a - b$ with tolerance $\tau_{\mathrm{H,G}}(d; M)$, then*

$$p_{\mathrm{rev}}(a, b; M) := \mathbb{P}\big[X^a_{\mathcal{F}}(\mathbf{m}) < X^b_{\mathcal{F}}(\mathbf{m})\big] = \mathbb{P}\big[X^{a-b}_{\mathcal{F}}(\mathbf{m}) < 0\big]\,,$$
$$\left| p_{\mathrm{rev}}(a, b; M) - \frac{1}{2} \right| \leq \tau_{\mathrm{H,G}}(a - b; M) + c_{\mathrm{G}}(\varepsilon). \tag{56}$$

Here $c_{\mathrm{G}}(\varepsilon)$ is the universal moment-replacement term defined in Eq. (21)*. For fixed finite h, define $\varepsilon_M := 2C_{\mathrm{F}}/((1-\rho)M\omega_{h-1})$ for all sufficiently large M. If the same shape calibration holds for the resulting fully-high partitions and $\tau_{\mathrm{H,G}}(a - b; M) \to 0$, then*

$$\lim_{M \to \infty} p_{\mathrm{rev}}(a, b; M) = \frac{1}{2}. \tag{57}$$

Fully-high membership controls moments but does not by itself make a fixed finite trigonometric sum Gaussian or guarantee chance-level reversal at a finite window. The theorem provides a calibrated error bound and an asymptotic conclusion only under the stated shape condition. The main theorem instead combines the local-response floor with a conservative finite-window reversal floor for the same ordered score margin.

Scope of the asymptotic premise. For fixed finite h and fixed nonzero d, the score satisfies $|X^d_{\mathcal{F}}(m)| \leq \sum_n |d_n|$ at every relative distance. Its limiting variance under the fully-high moment bounds is $U(d)^2/2 > 0$. Thus its standardized support remains bounded, while a Gaussian has positive tails beyond that bound. The all-threshold Gaussian-shape discrepancy therefore cannot vanish solely by increasing M in this fixed-dimensional setting. The vanishing-error implication above is retained as a conditional statement from the earlier analysis; the present paper uses the finite-window error bound and the finite-window context-scale theorem. The accepted small finite-window Gaussian approximation is consistent with this distinction.

B.4 Interpretation for p-RoPE

The main reversal-cutoff theorem and the reversal results above concern fully rotary RoPE. In a partial-RoPE variant (Barbero et al., 2025), the score of a fixed key with fixed content states can be written as

$$S^{\mathrm{p}}_K(m) = c_K + X^{b_{\mathcal{R}}}_{\mathcal{R}}(m),$$

where $\mathcal{R}$ indexes the retained rotating dimensions and the NoPE component c_K is independent of relative distance. This constant does not cancel against a fixed score barrier:

$$p^{\mathrm{p}}_{\mathrm{exc}}(K; S, M) = \mathbb{P}[X^{b_{\mathcal{R}}}_{\mathcal{R}}(\mathbf{m}) > S - c_K].$$

Depending on its sign, a sufficiently strong NoPE anchor can therefore relax or tighten the score-barrier requirement after the ceiling is recalibrated with the effective barrier $S - c_K$. Likewise, a favorable NoPE margin adds a nonzero DC term to the cross-key score difference, so the fully rotary one-half conclusion of Theorem 3 need not apply. Thus p-RoPE may avoid a particular score-barrier or fully rotary reversal cutoff; the magnitude of its NoPE share alone, without the sign of the induced score margin, is not sufficient.

The same constant cancels exactly from the positional interval comparison:

$$S^{\mathrm{p}}_K(m_{\mathrm{near}}) - S^{\mathrm{p}}_K(m_{\mathrm{far}}) = X^{b_{\mathcal{R}}}_{\mathcal{R}}(m_{\mathrm{near}}) - X^{b_{\mathcal{R}}}_{\mathcal{R}}(m_{\mathrm{far}}).$$

NoPE dimensions may be viewed as zero-frequency dimensions and hence never become certified high, but for the same reason they supply no changing signal with which to order the two intervals. If the retained rotating component is nonzero and satisfies all hypotheses of Theorem 2, including full certification and two-interval calibration, the same bound applies to that component. If no rotating component remains, the two sampled scores are identical and $p_{\mathrm{ord}} = 0$; this is positional indifference, not a reliable preference for the near interval. Hence an unrotated anchor can support score-barrier stability, but it cannot by itself provide long-distance positional discrimination.

C Proofs of the Theoretical Results

C.1 Proof of Proposition 2

Proposition 2 (RoPE score reduction)**.** *For a fully rotary RoPE head, as the real query and key blocks range over $\mathbb{R}^{2h}$, the class of fixed query–key scores and their finite real linear combinations as functions of relative distance is exactly $\{X_{\mathcal{F}}^{z} : z \in \mathbb{C}^h\}$; hence all subsequent full- and bandwise analyses reduce to studying X_J^z.*

Proof. The proof has two directions. We first rewrite each rotary plane in complex form and use real linearity; we then construct real rotary blocks for an arbitrary coefficient vector.

Let $q, k_+, k_- \in \mathbb{R}^{2h}$ denote one pre-RoPE query and two pre-RoPE keys. For each frequency n, group the corresponding real rotary coordinates into the blocks $q_n := (q_{n,1}, q_{n,2})^\top$ and $(k_\pm)_n := ((k_\pm)_{n,1}, (k_\pm)_{n,2})^\top$, and represent these blocks by

$$Q_n := q_{n,1} + iq_{n,2}, \qquad (K_\pm)_n := (k_\pm)_{n,1} + i(k_\pm)_{n,2}. \tag{58}$$

Let p_q and p_k be the query and key positions, set $m := p_q - p_k$, and write

$$R(\theta) := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{59}$$

For a position-independent attention-score scale $c_{\mathrm{att}} > 0$, define

$$S_\pm(m) := c_{\mathrm{att}} \sum_{n\in\mathcal{F}} \langle R(p_q\omega_n)q_n, R(p_k\omega_n)(k_\pm)_n\rangle\,. \tag{60}$$

The rotation matrices satisfy

$$R(\alpha)^\top R(\beta) = R(\beta - \alpha). \tag{61}$$

Consequently, the contribution of the nth rotary plane is

$$\begin{aligned} \langle R(p_q\omega_n)q_n, R(p_k\omega_n)(k_\pm)_n\rangle &= q_n^\top R\big((p_k - p_q)\omega_n\big)(k_\pm)_n \\ &= q_n^\top R(-m\omega_n)(k_\pm)_n. \end{aligned} \tag{62}$$

The same plane contribution admits the following complex representation:

$$Q_n\overline{(K_\pm)_n} = q_{n,1}(k_\pm)_{n,1} + q_{n,2}(k_\pm)_{n,2} + i\big(q_{n,2}(k_\pm)_{n,1} - q_{n,1}(k_\pm)_{n,2}\big). \tag{63}$$

Taking the real part after multiplication by $e^{im\omega_n}$ gives

$$\begin{aligned} \mathrm{Re}\Big(Q_n\overline{(K_\pm)_n}e^{im\omega_n}\Big) &= \big(q_{n,1}(k_\pm)_{n,1} + q_{n,2}(k_\pm)_{n,2}\big)\cos(m\omega_n) \\ &\quad + \big(q_{n,1}(k_\pm)_{n,2} - q_{n,2}(k_\pm)_{n,1}\big)\sin(m\omega_n) \\ &= q_n^\top R(-m\omega_n)(k_\pm)_n. \end{aligned} \tag{64}$$

Combining Eq. (60) and Eq. (64) yields

$$S_\pm(m) = \mathrm{Re}\sum_{n\in\mathcal{F}} \big(c_{\mathrm{att}}Q_n\overline{(K_\pm)_n}\big)e^{im\omega_n} = X_{\mathcal{F}}^{a_\pm}(m), \tag{65}$$

$$(a_+)_n := c_{\mathrm{att}}Q_n\overline{(K_+)_n}, \qquad (a_-)_n := c_{\mathrm{att}}Q_n\overline{(K_-)_n}. \tag{66}$$

Thus every fixed score has the required form. Real linearity in the coefficient vector shows that every finite real linear combination has the same form. In particular, setting

$$a_n := (a_+)_n, \qquad b_n := (a_-)_n, \qquad d_n := a_n - b_n \tag{67}$$

gives

$$D(m) := S_+(m) - S_-(m) = X_{\mathcal{F}}^{d}(m). \tag{68}$$

Conversely, fix any $z \in \mathbb{C}^h$. Choosing $Q_n = 1$ and $(K_+)_n = \overline{z_n/c_{\mathrm{att}}}$ for every n gives $c_{\mathrm{att}}Q_n\overline{(K_+)_n} = z_n$. Each chosen complex coordinate corresponds to a real two-dimensional rotary block, so this constructs real $q, k_+ \in \mathbb{R}^{2h}$ that realize $X_{\mathcal{F}}^{z}$. □

C.2 Finite-Window Fourier-Frame Bound

For $\alpha \in \mathbb{R}$, define the normalized finite-sum kernel

$$K_M(\alpha) := \frac{1}{M} \sum_{m=0}^{M-1} e^{im\alpha}. \tag{69}$$

For a finite frequency set $J \subseteq \mathcal{F}$, define

$$\Lambda_J^0 := \{0\} \cup \{+\omega_n : n \in J\} \cup \{-\omega_n : n \in J\}, \qquad \delta_J := \min_{\lambda \neq \lambda' \in \Lambda_J^0} \operatorname{dist}(\lambda - \lambda', 2\pi\mathbb{Z}), \tag{70}$$

and the finite-window Gram matrix and signed-frame defect

$$\mathcal{G}_J(M) := \left[K_M(\lambda' - \lambda)\right]_{\lambda,\lambda' \in \Lambda_J^0}, \qquad \mathfrak{d}_J(M) := \left\|\mathcal{G}_J(M) - I\right\|_{\mathrm{op}}. \tag{71}$$

Lemma C.1 (Finite-window Fourier-frame bound)**.** *If $\delta_J > 0$, then*

$$\mathfrak{d}_J(M) \le \frac{C_{\mathrm{F}}}{M\delta_J}. \tag{72}$$

In particular, $M\delta_J \ge 2C_{\mathrm{F}}/\varepsilon$ implies $\mathfrak{d}_J(M) \le \varepsilon/2$.

Proof. We decompose the off-diagonal kernel into two cosecant forms, apply the separated-point Hilbert inequality to each form, and then normalize by the window length.

The finite geometric-sum kernel satisfies

$$\sum_{m=0}^{M-1} e^{im(\lambda-\lambda')} = \frac{e^{i(M-1/2)(\lambda-\lambda')} - e^{-i(\lambda-\lambda')/2}}{2i\sin((\lambda-\lambda')/2)}. \tag{73}$$

In the off-diagonal quadratic form, the two numerator terms decompose into cosecant forms whose coefficient vectors differ only by diagonal unitary modulation. Apply the circular Hilbert inequality of Montgomery & Vaughan (1974, Theorem 1, Eq. (1.2)) with $x_r = \lambda_r/(2\pi)$ and modulo-one spacing $\delta_J/(2\pi)$. It bounds each term, including its factor $1/2$, by π/δ_J in operator norm. Dividing by M and applying the triangle inequality gives

$$\mathfrak{d}_J(M) \le \frac{2\pi}{M\delta_J}. \tag{74}$$

Thus the lemma holds with the universal choice $C_{\mathrm{F}} := 2\pi$. The stated sufficient condition for the defect bound follows by substitution. □

C.3 Proof of Proposition 1 and Lemma A.1

Proposition 1 (Analytical moment guarantees under finite windows). *For any tolerance $\varepsilon \in (0, 1/2)$ and certified window M where H is nonempty, the high-frequency attention score $S_H(m) := \Re\sum_{n\in H} z_n e^{im\omega_n}$ and its coefficient norm $U_H(z) := (\sum_{n\in H}|z_n|^2)^{1/2}$ over uniformly distributed relative distances $m \sim \mathrm{Unif}\{0,\dots,M-1\}$ satisfy*

$$|\mathbb{E}_m[S_H(m)]| \le \varepsilon U_H(z) \qquad \text{and} \qquad \left|\mathrm{Var}_m(S_H(m)) - \frac{1}{2}U_H(z)^2\right| \le \varepsilon U_H(z)^2. \tag{5}$$

Lemma A.1 (Finite-window moments and conditional Gaussian calibration)**.** *Fix $0 < \varepsilon < 1/2$, a certified window $M \ge \Gamma_\varepsilon(\rho)$, and fixed coefficients z with $U_H(z) > 0$. Use $H = \mathcal{H}_\varepsilon(M)$ from* Eq. (12) *and the uniformly sampled relative distance in* Eq. (13)*. For its random score Y_H^z, write*

$$\mu_H := \langle X_H^z \rangle_M, \qquad \sigma_H^2 := \mathrm{Var}_M(X_H^z). \tag{17}$$

Then the following moment bounds hold:

$$|\mu_H| \le \varepsilon U_H(z), \qquad \left|\sigma_H^2 - \tfrac{1}{2}U_H(z)^2\right| \le \varepsilon U_H(z)^2. \tag{18}$$

Because $0 < \varepsilon < 1/2$, Eq. (18) *implies* $\sigma_H^2 \geq (1/2 - \varepsilon)U_H(z)^2 > 0$ *whenever* $U_H(z) > 0$. *The standardization below is therefore well defined. The remaining, non-deterministic source of approximation error is the Gaussian-shape discrepancy of the standardized window law,*

$$\Delta_{\mathrm{H,G}}(z; M) := \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left[\frac{Y_H^z - \mu_H}{\sigma_H} \leq x \right] - \Phi_{\mathrm{std}}(x) \right| \leq \tau_{\mathrm{H,G}}. \tag{19}$$

Under this condition, the exact form of the Gaussian conclusion is

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}[Y_H^z \leq x] - \Phi_{\mathrm{std}}\left(\frac{\sqrt{2}\, x}{U_H(z)} \right) \right| \leq \tau_{\mathrm{H,G}} + c_{\mathrm{G}}(\varepsilon), \tag{20}$$

where the universal moment-replacement term is

$$c_{\mathrm{G}}(\varepsilon) := \frac{\varepsilon}{\sqrt{\pi}} + \frac{\log((1-2\varepsilon)^{-1})}{2\sqrt{2\pi e}} = O(\varepsilon) \qquad (\varepsilon \to 0). \tag{21}$$

We use the convention $\mathcal{N}(\mu, \sigma^2)$*: the target variance is* $U_H(z)^2/2$*, equivalently the target standard deviation is* $U_H(z)/\sqrt{2}$.

Proof. By the exact formulation in Appendix A.3, it is enough to establish the two moment inequalities in Eq. (18) and then transfer a $\tau_{\mathrm{H,G}}$ Gaussian-shape bound to Eq. (20). The proof proceeds in four stages. We first certify separation of the signed frequencies and convert that separation into blockwise Gram-matrix control. We then extract the mean and variance bounds and finally transfer the conditional Gaussian approximation from the exact moments to their calibrated targets.

Use the finite-sum kernel, signed-frequency spacing, and Gram-matrix defect defined in Appendix C.2.

We first verify the required signed-frequency separation. Because $M \geq \Gamma_\varepsilon(\rho)$, the certified-high set is nonempty. Write $H = \{0, \ldots, v\}$. To lower-bound the signed wrapped spacing, consider the four possible pair types. Distances from the signed frequencies to zero are at least ρ^v; same-sign distances are at least $(1-\rho)\rho^v$; opposite-sign distances are at least $2\rho^v$; and the wrap-around distance is larger than $2\pi - 2$. These cases cover every pair of distinct signed frequencies. Hence

$$\delta_H \geq (1-\rho)\rho^v. \tag{75}$$

Since $v \in H$,

$$M\delta_H \geq M(1-\rho)\rho^v \geq (1-\rho)\Gamma_\varepsilon(\rho) = \frac{2C_{\mathrm{F}}}{\varepsilon}. \tag{76}$$

Lemma C.1 therefore gives $\mathfrak{d}_H(M) \leq \varepsilon/2$.

To read off the individual blocks, define the vector g_H and the matrices G_H^- and G_H^+ by

$$(g_H)_n := K_M(\omega_n), \tag{77}$$

$$(G_H^-)_{n,n'} := K_M(\omega_{n'} - \omega_n), \qquad (G_H^+)_{n,n'} := K_M(\omega_n + \omega_{n'}). \tag{78}$$

Order Λ_H^0 as $(0, +H, -H)$. Since $K_M(-\alpha) = \overline{K_M(\alpha)}$, the complete centered Gram matrix is

$$\mathcal{G}_H(M) - I = \begin{bmatrix} 0 & g_H^\top & \overline{g_H^*} \\ \overline{g_H} & G_H^- - I & \overline{G_H^+} \\ g_H & G_H^+ & \overline{G_H^-} - I \end{bmatrix}. \tag{79}$$

Every displayed block is a compression of the full centered matrix. Therefore

$$\|g_H\|_2 \leq \mathfrak{d}_H(M), \qquad \left\|G_H^- - I\right\|_{\mathrm{op}} \leq \mathfrak{d}_H(M), \qquad \left\|G_H^+\right\|_{\mathrm{op}} \leq \mathfrak{d}_H(M). \tag{80}$$

We now convert these block bounds into moment estimates. Let $Z_H^z(m) := \sum_{n \in H} z_n e^{im\omega_n}$, so $X_H^z = \mathrm{Re}\, Z_H^z$. The first row of Eq. (79) gives

$$\langle X_H^z \rangle_M = \mathrm{Re}(g_H^\top z_H), \qquad |\langle X_H^z \rangle_M| \leq \mathfrak{d}_H(M) U_H(z). \tag{81}$$

Using $(\operatorname{Re} Z)^2 = \frac{1}{2}|Z|^2 + \frac{1}{2}\operatorname{Re}(Z^2)$,

$$\left\langle (X_H^z)^2 \right\rangle_M = \frac{1}{2} z_H^* G_H^- z_H + \frac{1}{2} \operatorname{Re}\left(z_H^\top G_H^+ z_H\right). \tag{82}$$

The block bounds imply

$$\left| \left\langle (X_H^z)^2 \right\rangle_M - \frac{1}{2} U_H(z)^2 \right| \le \mathfrak{d}_H(M) U_H(z)^2. \tag{83}$$

After subtracting the squared mean,

$$\left| \operatorname{Var}_M(X_H^z) - \frac{1}{2} U_H(z)^2 \right| \le \left(\mathfrak{d}_H(M) + \mathfrak{d}_H(M)^2\right) U_H(z)^2. \tag{84}$$

Finally,

$$\mathfrak{d}_H(M) + \mathfrak{d}_H(M)^2 \le \frac{\varepsilon}{2} + \frac{\varepsilon^2}{4} \le \varepsilon. \tag{85}$$

Substitution into Eq. (81) and Eq. (84) proves Eq. (18).

It remains to prove the conditional Gaussian conclusion. Assume $U_H(z) > 0$ and set $s_0 := U_H(z)/\sqrt{2}$. The variance bound in Eq. (18) gives

$$\sigma_H^2 \ge \left(\frac{1}{2} - \varepsilon\right) U_H(z)^2 > 0, \qquad \sqrt{1-2\varepsilon} \le \frac{\sigma_H}{s_0} \le \sqrt{1+2\varepsilon}. \tag{86}$$

Thus the standardized variable in Eq. (19) is well defined, and that condition implies

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}[Y_H^z \le x] - \Phi_{\mathrm{std}}\left(\frac{x - \mu_H}{\sigma_H}\right) \right| \le \tau_{\mathrm{H,G}}. \tag{87}$$

We next replace the exact moments by their calibrated targets. Differentiating $\Phi_{\mathrm{std}}((x-\mu)/s)$ with respect to $\log s$ shows that its absolute derivative is at most $1/\sqrt{2\pi e}$, uniformly in x and μ. Hence Eq. (86) yields

$$\sup_{x \in \mathbb{R}} \left| \Phi_{\mathrm{std}}\left(\frac{x - \mu_H}{\sigma_H}\right) - \Phi_{\mathrm{std}}\left(\frac{x - \mu_H}{s_0}\right) \right| \le \frac{\log((1-2\varepsilon)^{-1})}{2\sqrt{2\pi e}}. \tag{88}$$

Since the standard Gaussian CDF is $1/\sqrt{2\pi}$-Lipschitz, the mean bound in Eq. (18) also gives

$$\sup_{x \in \mathbb{R}} \left| \Phi_{\mathrm{std}}\left(\frac{x - \mu_H}{s_0}\right) - \Phi_{\mathrm{std}}\left(\frac{x}{s_0}\right) \right| \le \frac{|\mu_H|}{s_0\sqrt{2\pi}} \le \frac{\varepsilon}{\sqrt{\pi}}. \tag{89}$$

Combining Eq. (87), Eq. (88), and Eq. (89) by the triangle inequality proves Eq. (20). This completes the conditional Gaussian step. □

C.4 Strengthened Bandwise Calibration

The main text uses only the stated calibration. For completeness, the same frame argument also controls derivative variance and the score–derivative covariance on any band satisfying the stated separation condition.

Proposition 3 (Strengthened calibration for a separated band)**.** *Let $J \subseteq \mathcal{F}$ satisfy $M\delta_J \ge 2C_{\mathrm{F}}/\varepsilon$. Define*

$$E_0(J; z) := \sum_{n \in J} |z_n|^2, \qquad E_1(J; z) := \sum_{n \in J} \omega_n^2 |z_n|^2, \tag{90}$$

and

$$(X_J^z)'(m) := \operatorname{Re} \sum_{n \in J} i\omega_n z_n e^{im\omega_n}. \tag{91}$$

Then

$$|\langle X_J^z \rangle_M| \le \varepsilon \sqrt{E_0(J; z)}, \tag{92}$$

$$\left|\operatorname{Var}_M(X_J^z) - \tfrac{1}{2} E_0(J; z)\right| \le \varepsilon E_0(J; z), \tag{93}$$

$$\left|\operatorname{Var}_M((X_J^z)') - \tfrac{1}{2} E_1(J; z)\right| \le \varepsilon E_1(J; z), \tag{94}$$

$$|\operatorname{Cov}_M(X_J^z, (X_J^z)')| \le \varepsilon \sqrt{E_0(J; z) E_1(J; z)}. \tag{95}$$

Proof. Lemma C.1 gives $\mathfrak{d}_J(M) \le \varepsilon/2$. Repeating the block calculation in Appendix C.3 with H replaced by J proves the first two inequalities. For the derivative, set $w_n = i\omega_n z_n$. Then $X_J^w = (X_J^z)'$ and $\|w_J\|_2^2 = E_1(J; z)$, so the same variance calculation applies.

For the covariance, the bilinear version of Eq. (82) is

$$\langle X_J^z X_J^w \rangle_M = \frac{1}{2} \mathrm{Re}\big(z_J^* G_J^- w_J + z_J^\top G_J^+ w_J\big). \tag{96}$$

The identity contribution vanishes because $\mathrm{Re}(z_J^* w_J) = 0$. Eq. (80) and Eq. (81), with H replaced by J, therefore give

$$|\mathrm{Cov}_M(X_J^z, X_J^w)| \le \big(\mathfrak{d}_J(M) + \mathfrak{d}_J(M)^2\big) \|z_J\|_2 \|w_J\|_2 . \tag{97}$$

Finally, $\mathfrak{d}_J + \mathfrak{d}_J^2 \le \varepsilon/2 + \varepsilon^2/4 \le \varepsilon$. □

C.5 Direct Finite-Sum Intuition

The next calculation is independent of the Fourier-frame inequality above. It is not needed for the main-text calibration result, but it makes the high-frequency centering mechanism explicit.

Proposition 4 (Direct centering on the geometric grid)**.** *For a contiguous band $J = [u, v]$ of size $s = v - u + 1$,*

$$|\langle X_J^z \rangle_M| \le \frac{\pi}{M\rho^v} \sqrt{\frac{1-\rho^{2s}}{1-\rho^2}} U_J(z). \tag{98}$$

Proof. For $0 < \alpha \le 1$,

$$K_M(\alpha) = e^{i(M-1)\alpha/2} \frac{\sin(M\alpha/2)}{M \sin(\alpha/2)}, \qquad |K_M(\alpha)| \le \frac{\pi}{M\alpha}. \tag{99}$$

Therefore Cauchy–Schwarz gives

$$|\langle X_J^z \rangle_M| \le \frac{\pi}{M} \left(\sum_{n=u}^{v} \rho^{-2n} \right)^{1/2} U_J(z) \tag{100}$$

$$= \frac{\pi}{M\rho^v} \sqrt{\frac{1-\rho^{2s}}{1-\rho^2}} U_J(z). \tag{101}$$

□

C.6 Positional Response Bounds and Refinements

The proofs and coefficient-specific refinements in this section use the following appendix-only shorthand:

$$A_J(z) := \sum_{n \in J} |z_n| , \tag{102}$$

$$\Delta_1 X_J^z(t) := X_J^z(t+1) - X_J^z(t), \tag{103}$$

$$g_n := \left| e^{i\omega_n} - 1 \right| . \tag{104}$$

C.6.1 Auxiliary Non-High-Only Response Bound

Lemma C.2 (Non-high-only positional response has linear share cost)**.** *For every certified M and every nonzero z,*

$$\sup_t |\Delta_1 X_N^z(t)| \le \frac{\Gamma_\varepsilon(\rho)}{M} A_N(z; M). \tag{105}$$

Consequently, for every $\zeta > 0$, if $\sup_t |\Delta_1 X_N^z(t)| / U(z) \ge \zeta$, then

$$r_N(z; M) \ge \frac{\zeta M}{\Gamma_\varepsilon(\rho)\sqrt{h}}. \tag{106}$$

Proof. For every $n \in N$, the partition gives $\omega_n < \Gamma_\varepsilon(\rho)/M$. Hence

$$|\Delta_1 X_N^z(t)| \le \sum_{n\in N} g_n\,|z_n| \tag{107}$$

$$\le \sum_{n\in N} \omega_n\,|z_n| \tag{108}$$

$$\le \frac{\Gamma_\varepsilon(\rho)}{M} A_N(z;M), \tag{109}$$

which proves Eq. (105). Cauchy–Schwarz and $|N| \le h$ give

$$A_N(z;M) \le \sqrt{h}\,U_N(z) = \sqrt{h}\,U(z) r_N(z;M). \tag{110}$$

Therefore a normalized non-high response of size at least ζ implies

$$\zeta \le \frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{M} r_N(z;M), \tag{111}$$

which proves Eq. (106). □

The gain multiplier in this auxiliary envelope improves on the universal pointwise bound throughout the certified regime. Indeed, $M \ge \Gamma_\varepsilon(\rho)$ implies

$$0 < \frac{\Gamma_\varepsilon(\rho)}{M} \le 1 < 2, \tag{112}$$

where 2 is the universal bound on $\left|e^{i\omega} - 1\right|$. For the Qwen grid used in our experiments, $B = 10^6$ and $2h = 128$, so $\rho = (10^6)^{-1/64} \approx 0.805842$. With the fixed experimental choice $\varepsilon = 10^{-2}$, this gives

$$\Gamma_{10^{-2}}(\rho) \approx 6472.25, \qquad \frac{\Gamma_{10^{-2}}(\rho)}{32768} \approx 0.1975. \tag{113}$$

Thus the experimental window is certified, and its non-high gain coefficient is well below the trivial response coefficient.

C.6.2 Proof of Lemma A.2

Lemma A.2 (Direct scale-to-share response envelope)**.** *For every certified integer $M \ge 2$ and every $a \ne 0$,*

$$\max_{0\le m\le M-2} \frac{|X_{\mathcal{F}}^a(m+1) - X_{\mathcal{F}}^a(m)|}{U(a)} \le \underbrace{\frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{M} r_N(a;M)}_{\substack{\text{non-high band:}\\ \text{small rotations}}} + \underbrace{R_{\mathcal{F}}\, r_H(a;M)}_{\substack{\text{certified-high band:}\\ \text{coefficient norm}}} \le \frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{M} + R_{\mathcal{F}}\, r_H(a;M). \tag{22}$$

where $R_{\mathcal{F}} := \left(\sum_{n\in\mathcal{F}} \left|e^{i\omega_n} - 1\right|^2\right)^{1/2} > 0$ is the fixed full-grid adjacent-gain factor and is independent of M.

Proof. For every relative distance t, split the score difference over the non-high and certified-high bands. The triangle inequality gives

$$|\Delta_1 X_{\mathcal{F}}^a(t)| \le |\Delta_1 X_N^a(t)| + |\Delta_1 X_H^a(t)|\,. \tag{114}$$

Lemma C.2 and Cauchy–Schwarz imply

$$|\Delta_1 X_N^a(t)| \le \frac{\Gamma_\varepsilon(\rho)}{M} A_N(a;M) \le \frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{M} U_N(a). \tag{115}$$

For the certified-high band, another application of Cauchy–Schwarz gives

$$|\Delta_1 X_H^a(t)| \le \sum_{n\in H} g_n\,|a_n| \le \left(\sum_{n\in H} g_n^2\right)^{1/2} U_H(a) \le R_{\mathcal{F}} U_H(a). \tag{116}$$

Substitute these two bounds into Eq. (114), divide by $U(a) > 0$, and take the maximum over the integer window. This proves the first line of Eq. (22); the second follows from $r_N(a;M) \le 1$. □

C.6.3 Proof of Corollary A.1

Corollary A.1 (Stable local-response floor on the certified-high norm share)**.** *For every certified integer $M \geq 2$, every $a \neq 0$, and every $\zeta > 0$, if the score $X^a_{\mathcal{F}}$ avoids the window-level failure in* Eq. (15) *at response level ζ, then*

$$r_H(a; M) \geq \underline{r}_{\mathrm{pos}}(M; \zeta) := \frac{\left[\zeta - \Gamma_\varepsilon(\rho)\sqrt{h}/M\right]_+}{R_{\mathcal{F}}}. \tag{23}$$

where $R_{\mathcal{F}}$ is the fixed factor in Lemma A.2. If the right-hand side exceeds one, the requested response is impossible.

Proof. Combining the response requirement with Lemma A.2 yields

$$R_{\mathcal{F}} r_H(a; M) \geq \zeta - \frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{M}. \tag{117}$$

Because $r_H(a; M) \geq 0$ and $R_{\mathcal{F}} > 0$, take the positive part and divide by $R_{\mathcal{F}}$ to prove Eq. (23). A lower bound above one contradicts $r_H \leq 1$. □

C.6.4 Exact-Gain Response Envelope

For this appendix only, abbreviate the §2 window diagnostic as

$$\widehat{\overline{G}}_M(a) := \max_{0 \leq m \leq M-2} \frac{|X^a_{\mathcal{F}}(m+1) - X^a_{\mathcal{F}}(m)|}{U(a)}. \tag{118}$$

We also define the continuous normalized envelope

$$\overline{\mathcal{G}}(a) := \sup_{t \in \mathbb{R}} \frac{|\Delta_1 X^a_{\mathcal{F}}(t)|}{U(a)}. \tag{119}$$

To state two sharper relaxations, one coefficient-specific and one depending only on band norms, set

$$R_H(M) := \left(\sum_{n \in H} g_n^2\right)^{1/2}, \qquad R_N(M) := \left(\sum_{n \in N} g_n^2\right)^{1/2},$$
$$R^2_{\mathcal{F}} = R_H(M)^2 + R_N(M)^2. \tag{120}$$

The full-grid factor $R_{\mathcal{F}}$ is defined in Lemma A.2 and is independent of the partition. For $a \neq 0$ with $U_H(a) > 0$, also define

$$w_N(a; M) := \frac{\sum_{n \in N} g_n |a_n|}{U(a)}, \tag{121}$$

$$\kappa_H(a; M) := \frac{\sum_{n \in H} g_n |a_n|}{U_H(a)}, \tag{122}$$

$$T(a) := w_N(a; M) + \kappa_H(a; M) r_H(a; M) = \frac{\sum_{n \in \mathcal{F}} g_n |a_n|}{U(a)}. \tag{123}$$

Proposition 5 (Band-norm relaxations of the exact-gain envelope)**.** *For every certified integer $M \geq 2$ and every positional vector $a \neq 0$ with $U_H(a) > 0$, let $r = r_H(a; M)$ and define*

$$F_M(r) := R_N(M)\sqrt{1 - r^2} + R_H(M) r. \tag{124}$$

Then

$$\widehat{\overline{G}}_M(a) \leq \overline{\mathcal{G}}(a) \leq T(a) = w_N(a; M) + \kappa_H(a; M) r \leq w_N(a; M) + R_H(M) r \leq F_M(r). \tag{125}$$

Proof. The integer-window maximum ranges over a subset of the continuous relative distances, so $\widehat{\overline{G}}_M(a) \leq \overline{\mathcal{G}}(a)$. For every t, the triangle inequality gives

$$\frac{|\Delta_1 X^a_{\mathcal{F}}(t)|}{U(a)} \leq \frac{\sum_{n \in \mathcal{F}} g_n |a_n|}{U(a)} = T(a), \tag{126}$$

and hence $\overline{\mathcal{G}}(a) \le T(a)$. The identity $T(a) = w_N(a;M) + \kappa_H(a;M)r$ is given by Eq. (123). Cauchy–Schwarz on the high band gives

$$\kappa_H(a;M) \le \left(\sum_{n\in H} g_n^2\right)^{1/2} = R_H(M), \tag{127}$$

and hence

$$T(a) \le w_N(a;M) + R_H(M)\frac{U_H(a)}{U(a)} = w_N(a;M) + R_H(M)r. \tag{128}$$

Applying Cauchy–Schwarz to the non-high part gives

$$w_N(a;M) \le R_N(M)\frac{U_N(a)}{U(a)} = R_N(M)\sqrt{1-r^2}. \tag{129}$$

Substitution proves Eq. (125). □

The middle quantity $T(a)$ is a coefficient-specific, phase-agnostic upper certificate obtained from the triangle inequality. The final function F_M is an upper envelope that retains only the two band norms. Neither is a prediction of $\mathbb{E}[\widehat{\overline{G}}_M \mid r_H]$. In particular,

$$F_M(0) = R_N(M), \qquad r_*(M) := \frac{R_H(M)}{R_{\mathcal{F}}}, \tag{130}$$

so the envelope need not vanish at zero certified-high norm share and is increasing only for $0 \le r < r_*(M)$.

Corollary C.1 (Sharper coefficient-shape floor)**.** *For every certified integer $M \ge 2$, every $a \ne 0$ with $U_H(a) > 0$, and every $\zeta > 0$, define*

$$f_{\mathrm{pos}}(a;M,\zeta) := \frac{[\zeta - w_N(a;M)]_+}{\kappa_H(a;M)}. \tag{131}$$

If $\widehat{\overline{G}}_M(a) \ge \zeta$, then

$$r_H(a;M) \ge f_{\mathrm{pos}}(a;M,\zeta). \tag{132}$$

If the right-hand side exceeds one, the requested response is impossible.

Proof. Let $r = r_H(a;M)$. The response condition and Eq. (125) give

$$\zeta \le w_N(a;M) + \kappa_H(a;M)r. \tag{133}$$

Since $U_H(a) > 0$ and every g_n is positive, $\kappa_H(a;M) > 0$. Rearranging proves Eq. (132). □

Corollary C.2 (Monotone coefficient-specific relaxation)**.** *Under the assumptions of Corollary C.1, define*

$$f^{\mathrm{c}}_{\mathrm{pos}}(a;M,\zeta) := \frac{[\zeta - w_N(a;M)]_+}{R_{\mathcal{F}}}. \tag{134}$$

If $\widehat{\overline{G}}_M(a) \ge \zeta$, then

$$r_H(a;M) \ge f_{\mathrm{pos}}(a;M,\zeta) \ge f^{\mathrm{c}}_{\mathrm{pos}}(a;M,\zeta) \ge \underline{r}_{\mathrm{pos}}(M;\zeta). \tag{135}$$

For fixed a and ζ, this relaxation is non-decreasing in M.

Proof. Cauchy–Schwarz gives $\kappa_H(a;M) \le R_H(M) \le R_{\mathcal{F}}$, and hence $f_{\mathrm{pos}}(a;M,\zeta) \ge f^{\mathrm{c}}_{\mathrm{pos}}(a;M,\zeta)$. The leading lower bound in the stated hierarchy follows from Corollary C.1. Moreover, $g_n \le \omega_n < \Gamma_\varepsilon(\rho)/M$ for every $n \in N$, so Cauchy–Schwarz gives

$$w_N(a;M) \le \frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{M}, \tag{136}$$

so $f^{\mathrm{c}}_{\mathrm{pos}}(a;M,\zeta) \ge \underline{r}_{\mathrm{pos}}(M;\zeta)$. As M increases, $N(M)$ shrinks. Therefore $w_N(a;M)$ is non-increasing, while $R_{\mathcal{F}}$ is independent of the partition. Hence $f^{\mathrm{c}}_{\mathrm{pos}}(a;M,\zeta)$ is non-decreasing. □

C.6.5 General Separations

For a separation $\ell > 0$, define

$$\Delta_\ell X_j^z(t) := X_j^z(t+\ell) - X_j^z(t), \tag{137}$$

$$\beta_N(M,\ell) := \max_{n\in N} \left|e^{i\ell\omega_n} - 1\right|, \qquad R_H(M,\ell) := \left(\sum_{n\in H} \left|e^{i\ell\omega_n} - 1\right|^2\right)^{1/2}. \tag{138}$$

Proposition 6 (Response bound at a general separation)**.** *For every t,*

$$|\Delta_\ell X_{\mathcal{F}}^z(t)| \le \beta_N(M,\ell) A_N(z;M) + R_H(M,\ell) U_H(z) \le \frac{\ell\Gamma_\varepsilon(\rho)}{M} A_N(z;M) + R_H(M,\ell) U_H(z). \tag{139}$$

Proof. The first inequality follows from the triangle inequality on N and Cauchy–Schwarz on H. For the second inequality, use $\left|e^{i\ell\omega_n} - 1\right| \le \ell\omega_n < \ell\Gamma_\varepsilon(\rho)/M$ on N. □

C.6.6 Coarse Constant Floor

The following coarser form is sometimes convenient when only uniform norm budgets are available.

Corollary C.3 (Constant positional-share floor)**.** *Assume $A_N(a;M) \le A_*$ and $U(a) \le U_*$. If*

$$M \ge \max\left\{\Gamma_\varepsilon(\rho), \frac{2\Gamma_\varepsilon(\rho)A_*}{\eta}\right\} \tag{140}$$

and $|\Delta_1 X_{\mathcal{F}}^a(t)| \ge \eta$, then

$$r_H(a;M) \ge \frac{\eta}{4\sqrt{h}\,U_*}. \tag{141}$$

Proof. Taking $\ell = 1$ in Eq. (139), the assumed lower bound on the window length makes the non-high term at most $\eta/2$. Thus $R_H(M)U_H(a) \ge \eta/2$. Divide by $U(a) \le U_*$ and use $R_H(M) \le 2\sqrt{h}$. □

C.7 Proof of Theorem 2

Theorem 2 (Near–far ordering approaches chance under full calibration)**.** *Fix a fully rotary head together with fixed query and key content states that induce a nonzero coefficient vector a. Independently draw m_{near} uniformly from $\{M+1,\dots,2M\}$ and m_{far} uniformly from $\{2M+1,\dots,3M\}$. Suppose the two interval score laws lie in the asymptotic fully-high calibration regime of Appendix B.1.3. Then*

$$p_{\mathrm{ord}}(a;M) := \mathbb{P}[X_{\mathcal{F}}^a(m_{\mathrm{far}}) > X_{\mathcal{F}}^a(m_{\mathrm{near}})] \longrightarrow \frac{1}{2} \qquad (M\to\infty). \tag{39}$$

Proof. We compare both translated-window laws with a common Gaussian reference law. We first replace the near-window CDF in the ordering probability, then replace the far-window CDF, and finally pass to the asymptotic regime.

Let Y_{near} and Y_{far} be the independent interval scores in Eq. (39), with CDFs F_{near} and F_{far}. Since both translated windows have length M, are fully high, and satisfy the stated Gaussian-shape calibration, Proposition 1 compares them with the same continuous Gaussian reference CDF

$$G_a(x) := \Phi_{\mathrm{std}}\left(\frac{\sqrt{2}\,x}{U(a)}\right). \tag{142}$$

Specifically,

$$\begin{aligned} \sup_x |F_{\mathrm{near}}(x) - G_a(x)| &\le \eta_{\mathrm{near}} := \tau_{\mathrm{near}}(a;M) + c_{\mathrm{G}}(\varepsilon), \\ \sup_x |F_{\mathrm{far}}(x) - G_a(x)| &\le \eta_{\mathrm{far}} := \tau_{\mathrm{far}}(a;M) + c_{\mathrm{G}}(\varepsilon). \end{aligned} \tag{143}$$

Using the left limit of the near-interval CDF, the strict ordering event is

$$p_{\mathrm{ord}}(a;M) = \int_{\mathbb{R}} F_{\mathrm{near}}(y^-)\, dF_{\mathrm{far}}(y). \tag{144}$$

Continuity of G_a and Eq. (143) give

$$\left| p_{\mathrm{ord}}(a;M) - \int_{\mathbb{R}} G_a(y)\, dF_{\mathrm{far}}(y) \right| \le \eta_{\mathrm{near}}. \tag{145}$$

If Z has CDF G_a and is independent of Y_{far}, then

$$\begin{aligned} \int_{\mathbb{R}} G_a(y)\, dF_{\mathrm{far}}(y) &= \mathbb{P}[Z < Y_{\mathrm{far}}] \\ &= \int_{\mathbb{R}} \big(1 - F_{\mathrm{far}}(z)\big)\, dG_a(z). \end{aligned} \tag{146}$$

The second inequality in Eq. (143) and $\int (1 - G_a)\, dG_a = 1/2$ therefore imply

$$\left| \int_{\mathbb{R}} G_a(y)\, dF_{\mathrm{far}}(y) - \frac{1}{2} \right| \le \eta_{\mathrm{far}}. \tag{147}$$

Combining the last two bounds gives

$$\left| p_{\mathrm{ord}}(a;M) - \frac{1}{2} \right| \le \tau_{\mathrm{ord}}(a;M,\varepsilon). \tag{148}$$

The left-CDF identity handles the strict event without a no-ties assumption.

Finally, $\varepsilon_M \to 0$. Taking $\varepsilon = \varepsilon_M$ in Eq. (148) and using $\tau_{\mathrm{ord}}(a;M,\varepsilon_M) \to 0$ proves Eq. (39). □

C.8 Proof of Lemma B.1

Lemma B.1 (Calibrated single-key score exceedance)**.** *For every certified M, nonzero b, and $S \in \mathbb{R}$ in the non-degenerate calibration regime of Appendix B.2.2,*

$$|p_{\mathrm{exc}}(b;S,M) - \widehat{p}_{\mathrm{exc}}(b;S,M)| \le \tau_{\mathrm{exc}}(b;S,M). \tag{46}$$

Proof. We first compare the exact and calibrated moments, then compare the true tail with the exact-moment Gaussian approximation using the shape discrepancy, and finally propagate the moment replacements through the Gaussian CDF.

Let $\mu_H := \langle X_H^b \rangle_M$ and $\sigma_H^2 := \mathrm{Var}_M(X_H^b)$. The exact sum $Y_b = Y_N + Y_H$ gives

$$\mu_b = \mu_N + \mu_H, \qquad \sigma_b^2 = \sigma_N^2 + \sigma_H^2 + 2\kappa_{NH}. \tag{149}$$

Proposition 1, applied with $z = b$, therefore yields

$$|\mu_b - \mu_N| \le \varepsilon U_H(b), \qquad \left|\sigma_b^2 - \widehat{\sigma}_b^2\right| \le \varepsilon U_H(b)^2. \tag{150}$$

The assumed variance floor implies $\sigma_b \ge s_-(b;M) > 0$.

By the definition in Eq. (52) and the identity $1 - \Phi_{\mathrm{std}}(x) = \Phi_{\mathrm{std}}(-x)$,

$$\left| p_{\mathrm{exc}}(b;S,M) - \Phi_{\mathrm{std}}\left(\frac{\mu_b - S}{\sigma_b}\right) \right| \le \Delta_{\mathrm{exc}}(b;M). \tag{151}$$

A no-ties assumption is unnecessary because $p_{\mathrm{exc}} = 1 - \mathbb{P}[Y_b \le S]$ uses exactly the closed-threshold CDF appearing in Eq. (52).

The moment bounds also imply

$$\begin{aligned} \left| \frac{\mu_b - S}{\sigma_b} - \frac{\mu_N - S}{\widehat{\sigma}_b} \right| &\le \frac{|\mu_b - \mu_N|}{\sigma_b} + |S - \mu_N| \left| \frac{1}{\sigma_b} - \frac{1}{\widehat{\sigma}_b} \right| \\ &\le \frac{\varepsilon U_H(b)}{s_-(b;M)} + \frac{|S - \mu_N|\, \varepsilon U_H(b)^2}{s_-(b;M)\widehat{\sigma}_b\big(s_-(b;M) + \widehat{\sigma}_b\big)}. \end{aligned} \tag{152}$$

Here we used $|\sigma_b - \widehat{\sigma}_b| = \left|\sigma_b^2 - \widehat{\sigma}_b^2\right| / (\sigma_b + \widehat{\sigma}_b)$. Since Φ_{std} is $1/\sqrt{2\pi}$-Lipschitz, combining Eq. (151) and Eq. (152) proves Eq. (46). □

C.9 Proof of Corollary B.1

Corollary B.1 (Small score-barrier exceedance imposes a certified-high norm-share ceiling)**.** *Under the hypotheses of Lemma B.1, assume* $U_N(b)U_H(b) > 0$ *and* $0 < \delta + \tau_{\mathrm{exc}}(b; S, M) < 1/2$*. Then*

$$p_{\mathrm{exc}}(b; S, M) \le \delta \quad \Longrightarrow \quad r_H(b; M) \le u_{\mathrm{exc}}(b; S, M, \delta) < 1. \tag{47}$$

Proof. The proof converts the probability constraint into a variance bound, rewrites that bound in the share coordinate, and then solves the resulting quadratic inequality.

If $p_{\mathrm{exc}}(b; S, M) \le \delta$, then Lemma B.1 gives

$$\widehat{p}_{\mathrm{exc}}(b; S, M) \le \delta + \tau_{\mathrm{exc}}(b; S, M) < \frac{1}{2}. \tag{153}$$

By monotonicity and symmetry of the standard Gaussian CDF,

$$S - \mu_N \ge \xi_{\mathrm{exc}} \widehat{\sigma}_b > 0, \qquad \frac{\widehat{\sigma}_b^2}{U_N(b)^2} \le \frac{\lambda_S^2}{\xi_{\mathrm{exc}}^2}. \tag{154}$$

In particular, $\lambda_S > 0$. Define

$$t := \frac{U_H(b)}{U_N(b)} = \frac{r_H(b; M)}{\sqrt{1 - r_H(b; M)^2}} > 0. \tag{155}$$

Substituting the normalized descriptors into Eq. (154) gives the quadratic condition

$$\frac{1}{2}t^2 + 2\gamma_{NH}t + \beta_N \le \frac{\lambda_S^2}{\xi_{\mathrm{exc}}^2}. \tag{156}$$

The actual $t > 0$ is feasible, so the discriminant in Eq. (55) is nonnegative and its upper root $\bar{t}_{\mathrm{exc}}$ satisfies $t \le \bar{t}_{\mathrm{exc}}$; in particular, $\bar{t}_{\mathrm{exc}} > 0$. The map $t \mapsto t/\sqrt{1+t^2}$ is strictly increasing on $[0, \infty)$, which proves Eq. (47) and also gives $u_{\mathrm{exc}} < 1$. □

C.10 Proof of Theorem 1

Theorem 1 (Precise context-length limit on joint reliability)**.** *Fix a fully rotary head with the geometric frequency grid in* Eq. (8)*, a tolerance* $0 < \varepsilon < 1/2$*, and a nonzero ordered margin* $D(m) = X_{\mathcal{F}}^{d}(m)$ *with* $D(0) > 0$*. Let* $\mathfrak{M}_{\mathrm{cal}}$ *be a finite nonempty set of certified integer windows* $M \ge 2$*, with the same coefficients* d *at every window. Fix* $0 \le \delta < 1/2$*,* $0 < \zeta < R_{\mathcal{F}}$*, and* $0 \le \overline{\tau}_{\mathrm{G}} < 1/2$*. Assume, for every* $M \in \mathfrak{M}_{\mathrm{cal}}$*, that* $\sigma_D(M) > 0$ *and*

$$\Delta_{\mathrm{rev}}(d; M) = \left| p_{\mathrm{rev}}(d; M) - \Phi_{\mathrm{std}}\left(-\frac{\mu_D(M)}{\sigma_D(M)} \right) \right| \le \overline{\tau}_{\mathrm{G}}.$$

Let $\underline{p}_{\mathrm{rev}}(M; \zeta, \overline{\tau}_{\mathrm{G}})$ *be the explicit bound in* Eq. (30)*. This bound is non-decreasing across* $\mathfrak{M}_{\mathrm{cal}}$*, and each calibrated window retaining the adjacent response* $\max_{0 \le m \le M-2} g_d(m) \ge \zeta$ *satisfies*

$$p_{\mathrm{rev}}(d; M) \ge \underline{p}_{\mathrm{rev}}(M; \zeta, \overline{\tau}_{\mathrm{G}}).$$

Write $M_- := \min \mathfrak{M}_{\mathrm{cal}}$ *and* $M_+ := \max \mathfrak{M}_{\mathrm{cal}}$*. If the reversal tolerance satisfies*

$$\underline{p}_{\mathrm{rev}}(M_-) \le \delta < \underline{p}_{\mathrm{rev}}(M_+), \tag{35}$$

define $M_{\max} := M_{\max}^{\mathrm{cert}}$ *by* Eq. (34)*. Then, for every* $M \in \mathfrak{M}_{\mathrm{cal}}$ *with* $M > M_{\max}$*,*

$$p_{\mathrm{rev}}(d; M) > \delta \qquad \text{or} \qquad \max_{0 \le m \le M-2} g_d(m) < \zeta.$$

Thus $M_{\max}$ *is an upper bound on joint score-level reliability within the calibrated family. A window at or below* $M_{\max}$ *remains a candidate; the bound alone does not establish joint reliability there.*

An explicit sufficient condition for the strict upper inequality in Eq. (35) *is* $M_+ \ge M_{\mathrm{suff}}$ *and* $0 \le \delta < \delta_\star$*, where*

$$\delta_\star := \left[\Phi_{\mathrm{std}}\left(-\frac{\varepsilon}{\sqrt{1/2 - \varepsilon}} \right) - \overline{\tau}_{\mathrm{G}} \right]_+, \tag{36}$$

$$\omega_{\min} := \rho^{h-1},$$
$$M_{\mathrm{suff}} := \max\left\{2, \left\lceil \frac{\Gamma_\varepsilon(\rho)}{\omega_{\min}} \right\rceil, \left\lfloor \frac{\Gamma_\varepsilon(\rho)\sqrt{h}}{\zeta} \right\rfloor + 1\right\}. \tag{37}$$

Indeed, $\underline{p}_{\mathrm{rev}}(M) \le \delta_\star < 1/2$ at every calibrated window, with equality for each such window $M \ge M_{\mathrm{suff}}$. The lower inequality $\underline{p}_{\mathrm{rev}}(M_-) \le \delta$ remains required.

Proof. Fix a calibrated window at which the same margin d retains normalized adjacent response ζ. Corollary A.1 applied to d gives

$$r := r_H(d; M) \ge \ell(M; \zeta). \tag{157}$$

Write $N = \mathcal{N}_\varepsilon(M)$ and $H = \mathcal{H}_\varepsilon(M)$. The high-band moment bounds in Eq. (18) imply

$$\left|\left\langle X_H^d \right\rangle_M\right| \le \varepsilon U_H(d), \qquad \sqrt{\mathrm{Var}_M(X_H^d)} \ge c_\varepsilon U_H(d). \tag{158}$$

For the non-high band, Cauchy–Schwarz gives

$$\left|\left\langle X_N^d \right\rangle_M\right| \le \sum_{n\in N} |d_n| \le \sqrt{k_N(M)}\, U_N(d), \qquad \sqrt{\mathrm{Var}_M(X_N^d)} \le \sqrt{k_N(M)}\, U_N(d). \tag{159}$$

The second inequality follows because centering cannot increase the root mean square and $\left|X_N^d(m)\right| \le \sum_{n\in N} |d_n|$ pointwise.

Since $U_H(d) = U(d)r$ and $U_N(d) = U(d)\sqrt{1-r^2}$, the triangle and reverse triangle inequalities in finite-window L_2 yield

$$|\mu_D(M)| \le U(d) u_M(r), \qquad \sigma_D(M) \ge U(d) v_M(r). \tag{160}$$

Let $g_M(s) := c_\varepsilon s - \sqrt{k_N(M)}\sqrt{1-s^2}$; this function is non-decreasing on $[0,1]$ and $v_M(s) = [g_M(s)]_+$. If $v_M(r) = 0$, then $r \ge \ell(M;\zeta)$ implies $v_M(\ell(M;\zeta)) = 0$, so both envelope values in Eq. (31) vanish and the desired bound follows from $p_{\mathrm{rev}}(d; M) \ge 0$. Otherwise, Eq. (25) and Eq. (160) give

$$\begin{aligned} p_{\mathrm{rev}}(d; M) &\ge \Phi_{\mathrm{std}}\left(-\frac{\mu_D(M)}{\sigma_D(M)}\right) - \overline{\tau}_{\mathrm{G}} \\ &\ge \Phi_{\mathrm{std}}\left(-\frac{u_M(r)}{v_M(r)}\right) - \overline{\tau}_{\mathrm{G}}. \end{aligned} \tag{161}$$

For the second line, explicitly, $-\mu_D/\sigma_D \ge -|\mu_D|/\sigma_D \ge -u_M(r)/v_M(r)$. Since a probability is nonnegative, taking the positive part of the last lower bound proves $p_{\mathrm{rev}}(d; M) \ge \psi_M(r; \overline{\tau}_{\mathrm{G}})$.

It remains to verify monotonicity. On the region $v_M(r) > 0$, set $x(r) := \sqrt{1-r^2}/r$. Then

$$\frac{u_M(r)}{v_M(r)} = \frac{\varepsilon + \sqrt{k_N(M)}x(r)}{c_\varepsilon - \sqrt{k_N(M)}x(r)}. \tag{162}$$

The right-hand side is non-decreasing in x, whereas $x(r)$ decreases with r. Thus $\psi_M(r; \overline{\tau}_{\mathrm{G}})$ is non-decreasing in r on the positive-v_M region. On the preceding region $v_M = 0$, it is identically zero. If $k_N(M) > 0$, the ratio diverges at the boundary and its Gaussian lower tail tends to zero. If $k_N(M) = 0$, the ratio equals $\varepsilon/c_\varepsilon$ for every $r > 0$, and the prescribed value $\psi_M(0; \overline{\tau}_{\mathrm{G}}) = 0$ also preserves monotonicity. Positive-part clipping therefore preserves global monotonicity. Combining this fact with Eq. (157) proves Eq. (31).

Finally, $\ell(M;\zeta)$ is non-decreasing in M, while $k_N(M)$ is non-increasing because the certified-high prefix expands with M. The same ratio in Eq. (162) depends on these quantities through $\sqrt{k_N(M)}\sqrt{1-\ell(M;\zeta)^2}/\ell(M;\zeta)$, which is non-increasing whenever $\ell(M;\zeta) > 0$. When $\ell(M;\zeta) = 0$ or $v_M(\ell(M;\zeta)) = 0$, the floor is zero. Therefore $\underline{p}_{\mathrm{rev}}(M; \zeta, \overline{\tau}_{\mathrm{G}})$ is non-decreasing across $\mathfrak{M}_{\mathrm{cal}}$. Its minimum and maximum over this finite family are attained at M_- and M_+, respectively. Thus Eq. (35) is equivalent to the sub-threshold and crossing sets both being nonempty. At and beyond the first scale $M_\dagger$ in Eq. (32), the lower bound exceeds δ, so any margin retaining response ζ has $p_{\mathrm{rev}}(d; M) > \delta$. Finiteness and monotonicity make the largest sub-threshold window

the grid point immediately preceding $M_\dagger$ and rule out every larger calibrated scale. This proves $M_{\max} = M^{\mathrm{cert}}_{\max}$ in Eq. (34).

To prove the explicit sufficient condition, observe that whenever $v_M(r) > 0$, Eq. (162) and $x(r) \geq 0$ give $u_M(r)/v_M(r) \geq \varepsilon/c_\varepsilon$. Monotonicity of the Gaussian CDF and positive-part clipping imply $\psi_M(r; \overline{\tau}_{\mathrm{G}}) \leq \delta_\star$. The same bound holds when $v_M(r) = 0$, since ψ_M then vanishes. Consequently $\underline{p}_{\mathrm{rev}}(M) \leq \delta_\star < 1/2$. For $M \geq M_{\mathrm{suff}}$, the definition in Eq. (37) gives $M\omega_{\min} \geq \Gamma_\varepsilon(\rho)$ and $M > \Gamma_\varepsilon(\rho)\sqrt{h}/\zeta$. Hence $k_N(M) = 0$ and $\ell(M;\zeta) > 0$, so

$$\frac{u_M(\ell(M;\zeta))}{v_M(\ell(M;\zeta))} = \frac{\varepsilon}{c_\varepsilon}, \qquad \underline{p}_{\mathrm{rev}}(M) = \delta_\star.$$

If $M_+ \geq M_{\mathrm{suff}}$ and $0 \leq \delta < \delta_\star$, this equality proves the strict upper inequality in Eq. (35). Its lower inequality ensures that the sub-threshold set is also nonempty, completing the proof. □

C.11 Proof of Theorem 3

Theorem 3 (Fully-high pairwise reversal approaches chance)**.** *For a fully rotary head, fix a query and an ordered pair of keys with coefficient vectors $a, b \in \mathbb{C}^h$ such that $d := a - b \neq 0$. For any $0 < \varepsilon < 1/2$ and any certified window satisfying $\mathcal{H}_\varepsilon(M) = \mathcal{F}$, draw $\mathbf{m} \sim \mathrm{Unif}\{0, \dots, M-1\}$. If the Gaussian-shape condition in* Eq. (19) *holds for $z = d = a - b$ with tolerance $\tau_{\mathrm{H,G}}(d; M)$, then*

$$\begin{aligned} p_{\mathrm{rev}}(a, b; M) &:= \mathbb{P}\big[X^a_{\mathcal{F}}(\mathbf{m}) < X^b_{\mathcal{F}}(\mathbf{m})\big] = \mathbb{P}\big[X^{a-b}_{\mathcal{F}}(\mathbf{m}) < 0\big], \\ \left| p_{\mathrm{rev}}(a, b; M) - \frac{1}{2} \right| &\leq \tau_{\mathrm{H,G}}(a - b; M) + c_{\mathrm{G}}(\varepsilon). \end{aligned} \tag{56}$$

Here $c_{\mathrm{G}}(\varepsilon)$ is the universal moment-replacement term defined in Eq. (21)*. For fixed finite h, define $\varepsilon_M := 2C_{\mathrm{F}}/((1-\rho)M\omega_{h-1})$ for all sufficiently large M. If the same shape calibration holds for the resulting fully-high partitions and $\tau_{\mathrm{H,G}}(a - b; M) \to 0$, then*

$$\lim_{M \to \infty} p_{\mathrm{rev}}(a, b; M) = \frac{1}{2}. \tag{57}$$

Proof. We first derive the finite-window reversal bound and then obtain the asymptotic limit by letting the calibration error vanish.

Fix $0 < \varepsilon < 1/2$ and a window for which $\mathcal{H}_\varepsilon(M) = \mathcal{F}$. Because $d = a - b \neq 0$,

$$U_H(d) = U(d) > 0, \qquad Y^d_H = X^d_{\mathcal{F}}(\mathbf{m}). \tag{163}$$

Under the Gaussian-shape calibration assumed in the theorem, Proposition 1, applied with $z = d$, gives

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}[X^d_{\mathcal{F}}(\mathbf{m}) \leq x] - \Phi_{\mathrm{std}}\left(\frac{\sqrt{2}\, x}{U(d)} \right) \right| \leq \tau_{\mathrm{H,G}}(d; M) + c_{\mathrm{G}}(\varepsilon). \tag{164}$$

Let $x_k = -1/k$. Then $x_k \uparrow 0$ and $\{X^d_{\mathcal{F}}(\mathbf{m}) < 0\} = \bigcup_{k \geq 1} \{X^d_{\mathcal{F}}(\mathbf{m}) \leq x_k\}$. Continuity from below of probability and continuity of Φ_{std} allow $k \to \infty$ in Eq. (164), which proves Eq. (56). This left-limit argument controls the strict reversal event without imposing a no-ties assumption.

For the asymptotic claim, the definition in the theorem gives

$$\Gamma_{\varepsilon_M}(\rho) = M\omega_{h-1}, \qquad \mathcal{H}_{\varepsilon_M}(M) = \mathcal{F}, \qquad \varepsilon_M \longrightarrow 0. \tag{165}$$

Applying the finite-window reversal bound with $\varepsilon = \varepsilon_M$ and using $\tau_{\mathrm{H,G}}(d; M) \to 0$ together with $c_{\mathrm{G}}(\varepsilon_M) \to 0$ proves Eq. (57). □

D Empirical Validation of the Theory

We examine the numerical consistency and tightness of the local-response bounds on fixed model coefficient vectors, then document the frequency partition and score-distribution evidence used in the main text. The local-response experiment measures the maximum normalized adjacent gap that appears in the theoretical certificate. The task-level Positional Score in Appendix G.1.1 averages these gaps over relative distances. All experiments below hold cached coefficient vectors fixed while varying relative distance over the stated windows.

D.1 Local Positional-Response Envelope and Local-Response Floor

We audit the sharper coefficient-specific refinement of the stable local-response envelope, recorded in Appendix C.6.4. This experiment uses the theorem-side choice $C_{\mathrm{F}} = 2\pi$ and $\varepsilon = 10^{-2}$ at $M = 32768$. We evaluate Qwen3-8B and Llama-3.1-8B using ten global queries, 200 sampled key tokens, and all 32 first-layer attention heads. Thus each model contributes 64,000 query–head–key coefficient vectors. Qwen3 has the geometric RoPE grid in Eq. (8); the certified partition has $|H| = 8$ and $|N| = 56$, so this model lies directly within the assumptions of the main theory. For Llama-3.1, we select the largest prefix satisfying the realized signed-frequency separation condition in the Fourier-frame argument, which gives $|H| = 9$ and $|N| = 55$. Since Llama-3.1 uses scaled, non-geometric RoPE frequencies, we report its row as a stress test outside the geometric-grid assumptions. This subsection audits the local result in Appendix A.4.1; it does not test the two-interval calibration in the near–far theorem of Appendix B.1.2.

For each coefficient vector a_i, we use the observable gap ζ_i and its pointwise certificate $T(a_i)$ from Proposition 5. The proposition and Corollary C.1 require, respectively,

$$\zeta_i \le T(a_i), \qquad f_{\mathrm{pos}}(a_i; M, \zeta_i) \le r_H(a_i; M). \tag{166}$$

Figure 7 reports all 128,000 vectors. Every point satisfies both inequalities in Eq. (166). In the left column, the median tightness ratios $\zeta_i / T(a_i)$ are 0.829 for Qwen3 and 0.788 for Llama-3.1. The middle column shows a strong monotone association between the certified-high norm share and both responses: the Spearman correlations of r_H with $(\zeta_i, T(a_i))$ are $(0.975, 0.969)$ for Qwen3 and $(0.988, 0.989)$ for Llama-3.1. This panel describes the sampled coefficient vectors, because $T(a_i)$ retains the point-specific quantities $w_N(a_i; M)$ and $\kappa_H(a_i; M)$.

In the right column, the median ratios $f_{\mathrm{pos}}(a_i; M, \zeta_i)/r_H(a_i; M)$ are 0.782 for Qwen3 and 0.743 for Llama-3.1. Here ζ_i is the observed maximum gap of the same coefficient vector used to compute f_{pos}. This is a same-sample consistency and tightness check. The proximity of the lower bound to the equality line quantifies its empirical tightness.

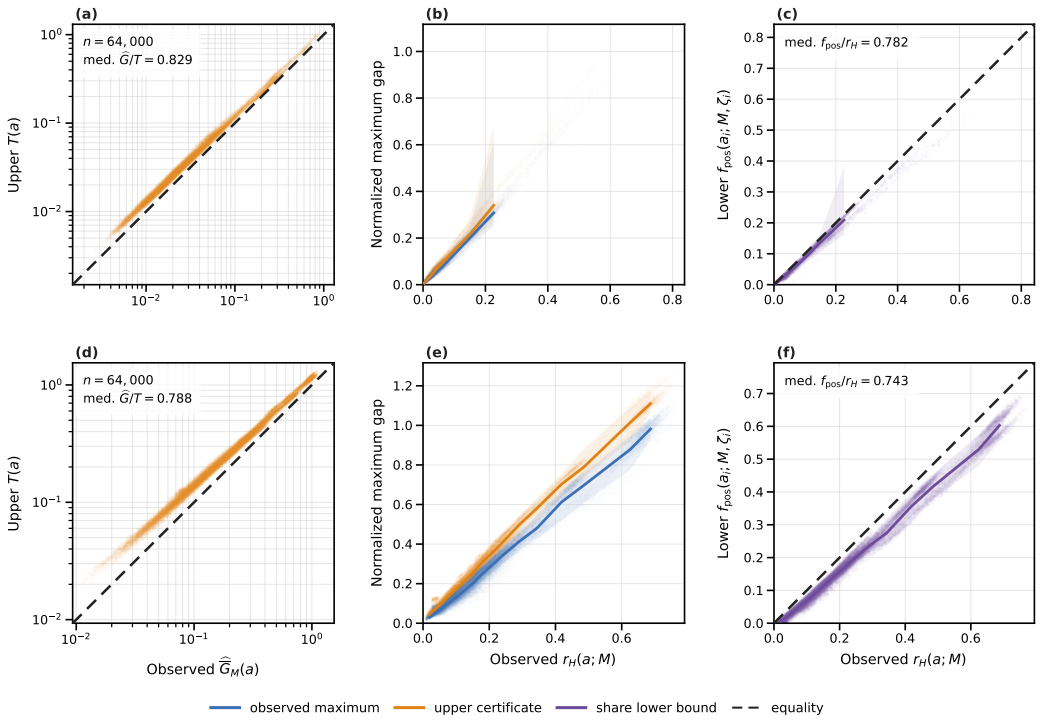


Figure 7: Pointwise consistency and tightness audit of the sharper coefficient-specific local positional-response envelope and local-response floor from the appendix. Panels (a)–(c) form the Qwen3-8B row and panels (d)–(f) form the Llama-3.1-8B row; every point is one first-layer query–head–key coefficient vector. Left: the observed integer-window maximum gap ζ_i against the coefficient-specific upper certificate $T(a_i)$ on logarithmic axes. Middle: the observed gap (blue) and upper certificate (orange) against $r_H(a_i; M)$; solid curves and shaded regions are medians and 10–90% ranges over 32 equal-count r_H bins, not an r_H-only predictor. Right: the pointwise lower bound $f_{\mathrm{pos}}(a_i; M, \zeta_i)$ against the observed certified-high norm share. Dashed diagonals indicate equality in the left and right columns. The Qwen3 row is an in-assumption geometric-grid audit, whereas the non-geometric Llama row is an out-of-assumption stress test. The right column uses the same point to set ζ_i and therefore measures consistency and tightness rather than out-of-sample predictive accuracy.

D.2 Experimental Details for Figure 3

This subsection gives the experimental protocols for §3.1 and Figure 3.

D.2.1 Empirical Choice of the Frequency Split

For practical frequency selection, we use an empirical value $C_{\mathrm{F}}^{\mathrm{emp}}$ for the constant in the theoretical split criterion:

$$H_{\mathrm{op}}(M) = \left\{ n : M\omega_n \geq \frac{2C_{\mathrm{F}}^{\mathrm{emp}}}{(1-\rho)\varepsilon_{\mathrm{ref}}} \right\}, \qquad C_{\mathrm{F}}^{\mathrm{emp}} = 0.06, \quad \varepsilon_{\mathrm{ref}} = 0.1. \tag{167}$$

Here ω_n are the actual model frequencies, including Llama-3.1's native frequency scaling, and $\rho = B^{-1/h}$ is computed from the base grid. We choose $C_{\mathrm{F}}^{\mathrm{emp}} = 0.06$ based on empirical observations from the evaluated models and windows, and use this single value across both models and all three context lengths. The selected bands achieve randomness scores between 0.9936 and 0.9997 under the metric defined below, showing close agreement between measured and predicted noise amplitudes across all six settings.

The proof of Lemma C.1 uses $C_{\mathrm{F}} = 2\pi$ as a conservative sufficient constant for a uniform bound over arbitrary fixed coefficient vectors. This choice establishes the theoretical guarantee in Appendix A.3;

$C_{\mathrm{F}}^{\mathrm{emp}}$ adapts the split criterion to the observed model coefficients. The reference value $\varepsilon_{\mathrm{ref}}$ controls the empirical threshold, and we assess the resulting split through measured noise-amplitude agreement. The empirical rule also applies to Llama-3.1's scaled frequency grid, whose evaluation lies outside the geometric-grid assumptions of the theoretical guarantee. The intervention experiments retain their original parameter $c_{\mathrm{int}} = 0.05$, as documented in Appendix G.2.

For Figures 3b and 3c, we evaluate Qwen3-8B (Yang et al., 2025a) and Llama-3.1-8B (Dubey et al., 2024) at $M \in \{8192, 32768, 131072\}$. For each model, the same 40,960 first-layer query–pair–head coefficient vectors are used at all three window lengths. We hold these cached vectors fixed and evaluate their RoPE score margins over $m \in \{0, \dots, M-1\}$, using the paper's convention $m =$ query position $-$ key position and margin coefficients $d_n = Q_n\overline{(K_{a,n} - K_{b,n})}/\sqrt{d_{\mathrm{head}}}$ for keys a and b. Thus the 128k curves extend the relative-distance window of fixed representations; they require no 128k model forward pass. For every prefix $H_k = \{0, \dots, k-1\}$, we compute the aggregate variance ratio

$$\mathcal{R}_{\mathrm{var}}(H_k; M) := \frac{\sum_i \mathrm{Var}_M(X_{H_k}^{d_i})}{\sum_i U_{H_k}(d_i)^2/2}, \tag{168}$$

where i indexes the sampled query–pair–head points and Var_M uses the uniform distribution over the M integer distances. We then compare noise amplitudes through

$$a(H_k; M) = \sqrt{\mathcal{R}_{\mathrm{var}}(H_k; M)}, \qquad R(H_k; M) = \max\{0, 1 - |a(H_k; M) - 1|\}. \tag{169}$$

The amplitude ratio a is formed after aggregating the measured and predicted variances. The plotted randomness proxy $R \in [0, 1]$ equals one at exact amplitude agreement and decreases with the absolute relative amplitude error. For example, ratios $a = 0.9$ and $a = 1.1$ both give $R = 0.9$. Positive cross-frequency covariance can give $\mathcal{R}_{\mathrm{var}} > 1$; the transformation preserves this discrepancy as a decrease in R. We retain all resulting dips and recoveries. This metric measures noise-scale agreement and supplies no test of Gaussian shape or temporal independence.

In Figures 3b and 3c, the gray curves plot $R(H_k; M)$ as the number k of retained components increases. Orange circles mark the largest prefixes satisfying Eq. (167) for 8k, 32k, and 128k windows, from left to right. The main axes magnify the selected cutoffs and nearby changes; insets show all 64 prefixes, with orange boxes indicating the enlarged regions.

D.2.2 Score-Distribution Illustration

Figure 3a reuses the Qwen3-8B layer-0 projection bank. For each of its 10 query tokens and 32 query heads, we uniformly sample four distinct key tokens from the existing 200-token key bank (seed 20260923), giving 1,280 query–key–head combinations. These are individual attention scores; Figures 3b and 3c use score margins. For each fixed combination i, we evaluate $S_{H_{40}}^{(i)}(m)$ at every $m \in \{0, \dots, 32767\}$ and divide by $\sigma_i = U_{H_{40}}(z^{(i)})/\sqrt{2}$, where $H_k = \{0, \dots, k-1\}$. The gray bars average the resulting densities with equal weight. No empirical centering or variance fitting is used. The orange curve is $\mathcal{N}(0, 1)$ under this normalization. This illustration uses the 32k cutoff from Eq. (167) and the same cached projections, with no new model forward pass.

D.3 Additional Evidence on High-Frequency Scores

D.3.1 The Concentrated Qwen Frequency Behind the Noise-Amplitude Drop

The sharp drop in the inset of Figure 3b occurs when the retained prefix grows from 51 to 52 frequencies. The newly included frequency has zero-based index $n = 51$, so it is the 52nd complex coordinate pair, corresponding to real channels 51 and 115 in the split-half layout. On the same 40,960 margin-coefficient vectors d_i used in that panel, its aggregate coefficient-energy share is

$$\frac{\sum_i |d_{i,51}|^2}{\sum_i \sum_n |d_{i,n}|^2} = 92.3507\%. \tag{170}$$

This is the fraction of the summed squared coefficient norms, with each sample entering the sum once. It is distinct from a fraction of the measured score variance or a norm share before squaring.

At $M = 32768$, this frequency rotates by only $M\omega_{51} = 0.54225$ radians over the window. Its own finite-window variance is 4.4599% of the half-energy prediction. Including it multiplies the

total predicted variance by 16.1731, while the aggregate variance ratio $\mathcal{R}_{\text{var}}$ falls from 0.883744 to 0.086922 and the plotted randomness proxy R falls from 0.940077 to 0.294826. The slowly varying component therefore adds substantial coefficient energy without the variance expected from a fast oscillation. This accounts for the abrupt drop and illustrates the need to distinguish coefficient magnitude from frequency-dependent variation. The component lies outside all three operational high-frequency prefixes.

D.3.2 Representation Assumptions and Gaussian Score Shape

The moment guarantees in Proposition 1 allow arbitrary fixed coefficient magnitudes and phases. Relevant prior analyses use several different assumptions. Xu et al. (2024) derive their expected preference bound from independent identically distributed query and key coordinates with a common variance and a similar-key noise model. Their theorem does not require Gaussian coordinate distributions. Du et al. (2026) also fix the query and key vectors and sample relative distance. Their high-frequency analysis gives the near-zero mean and half-energy variance approximation; its normal approximation uses regular amplitudes without a dominant component and approximate independence across frequencies. Our finite-window moment bounds control cross-frequency terms explicitly and allow concentrated coefficient energy.

Deterministic analyses also appear in Su et al. (2024), Barbero et al. (2025), and Urrutia et al. (2026). In particular, the Gaussian query/key model in Proposition 3.2 of Barbero et al. (2025) is a counterexample to universal distance decay; their other constructions and semantic-attention results use fixed vectors. These results establish that fixed-vector analysis is already part of the literature. The present contribution is the finite-window moment control and the resulting frequency criterion. Gaussian score shape remains a separate approximation: the displayed pooled density supports it empirically, and the figure does not establish that every individual query–key combination has a Gaussian window distribution.

E THE 49 REFERENCE TASK SETTINGS

The diagnostic reference set contains 49 task/configuration settings from nine benchmark sources. A setting combines a task with a difficulty or input-length configuration. Table 1 lists every setting by grouping configurations of the same task. The set covers key and document retrieval, multi-hop question answering, text and list ordering, state tracking, and long-context reasoning. For example, the LIFBench adjacent-element retrieval task contributes three length settings (3k, 6k, and 13k), while RULER2 multi-key retrieval contributes four difficulty settings (basic, easy, medium, and hard). These 49 settings form the ranking reference set in Figure 4; each point in Figure 6 and each bar in Figure 5 corresponds to one setting.

Qwen3-8B and Llama-3.1-8B-Instruct use the same example IDs within each setting: 2,211 examples per model and 4,422 example–model records in total. We record query and key activations during standard evaluation of these examples. The resulting diagnostics supplement each task's behavioral accuracy with information about relative failure susceptibility.

Sources and configurations. The 12 MK, MV, and QA settings use the official NeMo Skills RULER2 suite.[2] We use its nominal 8k configuration. Each family has basic, easy, medium, and hard variants, with 100 examples per configuration. The two LongBench tasks use unmodified examples whose native chat inputs fit within 16,384 tokens for both models. L-Eval TopicRet contributes 150 question records from 50 conversations, with three topic-retrieval questions per conversation. FLenQA uses the books-padding, random-evidence configuration at context sizes 500 and 3,000. NoLiMa uses book 1 with the needle at 52% depth. The three LIFBench settings contain 12 examples each; the remaining 33 settings contain 25 examples each.

Length labels in the table retain the source benchmark's configuration names. For the diagnostic calculation, M is the complete input length under the evaluated model's tokenizer, including its prompt and chat template. The same length label can therefore yield different M across models. The counts below describe the diagnostic reference set; matched coverage for the intervention experiments is reported separately in Appendix H.

Table 1: The complete 49-setting diagnostic reference set. Each configuration in a row is a separate setting. K is the number of settings and N is the number of examples per setting and per model. Lengths are nominal benchmark configurations; the diagnostic uses each model's actual input length.

Task	What the model must do	Configurations	K	N
RULER2 (NeMo Skills)				
Multi-key (MK)	Retrieve by key; harder variants answer retrieved MMLU questions.	Basic, easy, medium, hard	4	100
Multi-value (MV)	Retrieve multiple values or select an ordered question for a key.	Basic, easy, medium, hard	4	100
Question answering	Retrieve supporting documents and answer HotpotQA questions.	Basic, easy, medium, hard	4	100
LongBench-v1 (Bai et al., 2024)				
Qasper	Answer questions about an academic paper.	Native chat $\leq$ 16k	1	25
MuSiQue	Combine evidence across documents for multi-hop answers.	Native chat $\leq$ 16k	1	25
L-Eval (An et al., 2024)				
TopicRet	Retrieve the first, second, or third conversation topic.	Native conversation	1	150
Ada-LEval (Wang et al., 2024a)				
TSort	Restore the original order of shuffled text fragments.	1k	1	25

Continued on the next page

[2] NeMo Skills RULER2 task definitions.

Table 1 (continued)

Task	What the model must do	Configurations	K	N
FLenQA (Levy et al., 2024)				
PIR	Infer a room property through a person–room relation.	500; 3,000	2	25
MonoRel	Infer a transitive conclusion from ordered relations.	500; 3,000	2	25
Simplified RuleTaker	Deduce a true/false conclusion from rules and facts.	500; 3,000	2	25
LongReason (Ling et al., 2025)				
Mixed reasoning	Solve reading, logical, and mathematical reasoning questions.	8k; 16k	2	25
BABILong (Kuratov et al., 2024)				
qa1: person location	Track a person's current location.	8k; 16k	2	25
qa2: object location	Track an object's current location.	8k; 16k	2	25
qa3: previous location	Recover an object's location before a specified event.	8k; 16k	2	25
qa4: directional relations	Infer a directional relation between two entities.	8k; 16k	2	25
qa5: giving events	Identify the giver, recipient, or object in a giving event.	8k; 16k	2	25
qa7: counting	Count the objects a person currently holds.	8k; 16k	2	25
qa8: lists and sets	List the objects a person currently holds.	8k; 16k	2	25
qa9: negation	Answer location questions with explicit negation.	8k; 16k	2	25
qa10: indefinite knowledge	Distinguish known, false, and uncertain locations.	8k; 16k	2	25
LIFBench (Wu et al., 2025b)				
List offset query	Return the item immediately before or after a target.	Official base 3k; 6k; 13k	3	12
NoLiMa (Modarressi et al., 2025)				
One-hop retrieval	Retrieve a fact through one implicit semantic association.	8k; 16k	2	25
Two-hop retrieval	Retrieve a fact through two implicit semantic associations.	8k; 16k	2	25
Total: 49 settings; 2,211 examples per model			**49**	

F Token and Head Selection

F.1 Token Sampling from the Evaluated Inputs

The current diagnostic reuses the recorded query and key activations for the 49 reference settings in Appendix E. Each example is tokenized with the evaluated model's native chat template. Character spans supplied by the task preparation identify the question and its context. Eligible token occurrences overlap the corresponding span, contain at least one letter or digit within that overlap, and have nonempty character offsets; special tokens are excluded. We select up to four query positions from the question and up to 16 key positions from the context. Every selected key precedes every selected query in the original input.

Selection uses a deterministic SHA-256 ordering without replacement. For the original 12 RULER2 settings, the ordering is keyed by the protocol identifier, example ID, sampling role, and token position. The remaining 37 settings also include the fixed seed 20260910. Separate sampling roles select queries, keys, and the ordering used to pair keys. The saved token positions remain fixed throughout the diagnostic calculations.

For the *Common-pair* branch, each query is combined with every selected key, giving up to 64 query–key pairs per example. For the *Common-triplet* branch, the selected keys are reordered with the key-pairing role and grouped into eight disjoint pairs; each query is combined with these pairs, giving up to 32 triplets with distinct keys. All 2,211 examples per model have 16 selected keys. Of these, 2,161 have four selected queries and yield 64 pairs and 32 triplets. The other 50 examples, from the two BABILong person-location settings, have three eligible queries and yield 48 pairs and 24 triplets. The Common branches use these sampled occurrences without task-target labels. The semantic calculation subsequently orients each triplet by its score ordering at zero relative distance, as defined in Appendix G.1.1.

F.2 Selecting a Fixed Set of Attention Heads

We select heads whose high-frequency norm share varies across the reference settings. For each layer–head unit h, we first average the Common-pair value r_H over valid pairs within each example, then average over examples within each setting. Let $\bar{r}_{t,h}$ be this mean for setting t. With all 49 settings weighted equally, the selection statistic is the sample variance

$$v_h = \frac{1}{48} \sum_{t=1}^{49} \left(\bar{r}_{t,h} - \frac{1}{49} \sum_{s=1}^{49} \bar{r}_{s,h} \right)^2 . \tag{171}$$

Each model is ranked separately by decreasing v_h. Ties are resolved by increasing layer-major head index; neither model has a tie at the reported selection boundaries. A fraction p retains $\lceil pH \rceil$ of the model's H heads.

Qwen3-8B has 36 layers and 32 query heads per layer, for 1,152 heads in total; its top 5% and top 10% sets contain 58 and 116 heads. Llama-3.1-8B-Instruct has 32 layers and 32 query heads per layer, for 1,024 heads; its corresponding sets contain 52 and 103 heads. Figure 8 shows the selection across layers and heads. The top-5% set is fixed across tasks and supplies both the Semantic and Positional Scores in the main diagnostic. The second intervention stage expands this same ranking to the top 10%, as detailed in Appendix H. This ranking describes variation within the chosen reference set, including its different input-length settings.

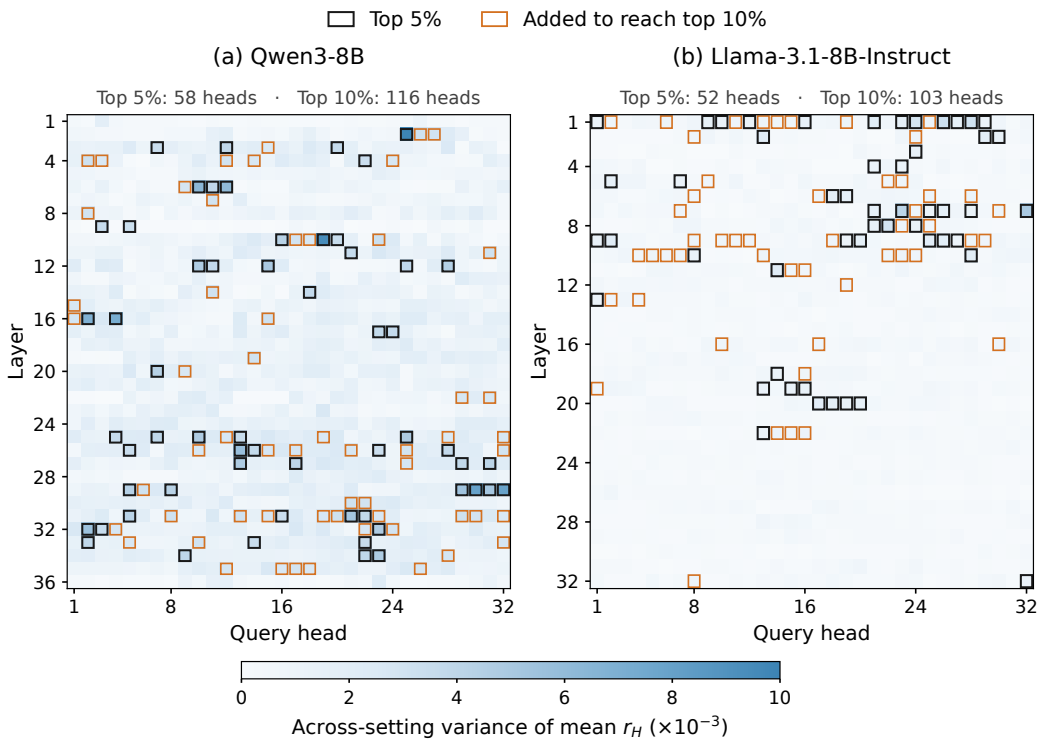


Figure 8: Head selection from the 49 reference settings. Each cell shows the sample variance of a head's setting-level mean Common-pair r_H, using a shared color scale across models. Black outlines identify the top-5% heads used by the main diagnostic and first intervention stage. Orange outlines identify the additional heads included in the top-10% set for the second stage. Layer and query-head indices start at one. All settings within a model use the same selected heads.

G Experimental Details

We conduct two experiments on Qwen3-8B and Llama-3.1-8B-Instruct: we compare diagnostic profiles across the 49 shared task settings, and we evaluate high-frequency interventions in each setting's assigned semantic or positional direction. Appendix E specifies the task data, and Appendix F describes the fixed token and head selections.

G.1 Experiment 1: Task and Cross-Model Diagnostic Profiles

Each model uses 2,211 examples across the same 49 settings. The primary diagnostic uses the fixed top-5% head collection selected from cross-task variation in the high-frequency norm share. All scores are computed from cached query and key states at each example's native input length and frequency configuration.

G.1.1 State Extraction and Score Definitions

For each task example, we use sampled query–key pairs and query–key–key triplets, together with a fixed collection of layer–head units. In each rotary plane, let Q_n and K_n denote the complex forms of the pre-RoPE query and key components. Their score coefficient is

$$z_n = c_{\mathrm{att}} Q_n \overline{K_n}, \qquad S(m) = c_0 + \mathrm{Re} \sum_n z_n e^{im\omega_n}, \tag{172}$$

where c_{att} is the model's attention-score scale and c_0 is any non-rotary score contribution (zero for fully rotary heads). The extraction preserves native query/key normalization, rotary-coordinate pairing, and the mapping between query heads and key/value heads. For partially rotary heads, the non-rotary contribution is retained as a constant in semantic margins; the norm used for positional response refers to the rotary coefficients.

We keep these content vectors fixed and vary only the current layer's RoPE phase over $m = 0, \ldots, M-1$. The current task protocol sets M to the example's input length. The score at zero relative distance supplies the reference comparison. These are virtual score evaluations and require no new forward pass once the states have been recorded.

Semantic Score. For each triplet, orient the margin (including any non-rotary constant) so that $D(0) > 0$, with $d = z_+ - z_-$ in this ordering, and compute the full finite-window mean μ_D and variance σ_D^2 analytically. Finite geometric sums give these moments, including the covariance between frequency components; Appendix G.1.2 provides the expressions. With Φ denoting the standard normal cumulative distribution function, the Gaussian estimate is

$$\widehat{p}_{\mathrm{rev}}(d; M) = \Phi\left(-\frac{\mu_D}{\sigma_D}\right), \qquad \sigma_D > 0. \tag{173}$$

We average these probabilities over valid triplets, then over the fixed head collection, and finally over task examples with equal example weights. The *Semantic Score* is one minus this task-level reversal rate. Higher values indicate more stable reference orderings. Reference ties are excluded and their counts are retained. For zero variance, we evaluate the deterministic strict-negative event directly.

The probability calculation uses the accepted Gaussian approximation and analytic moments. It avoids enumerating reversal events at every position. The high-frequency share provides the theoretical interpretation in §3; the practical estimate uses the full finite-window moments and coefficient vector. Its value may vary non-monotonically with the window even though the conservative theoretical lower bound is monotone. This distinction allows the tool to report the behavior of sampled coefficients while the theory explains a necessary constraint on their joint semantic and positional reliability.

Positional Score. For each query–key pair with $U(z) > 0$, define

$$P(z; M) := \frac{1}{M-1} \sum_{m=0}^{M-2} \frac{|S(m+1) - S(m)|}{U(z)}. \tag{174}$$

We first normalize each pair by its own coefficient norm, then average over pairs, the same fixed head collection, and task examples. The resulting *Positional Score* measures average normalized

adjacent response. Higher values indicate stronger local response. The score is computed directly from adjacent differences; a zero coefficient norm is recorded as undefined. This convention preserves the per-pair scale before aggregation.

The response condition in Theorem 1 applies to the pairwise margin $d = a - b$. Requiring the average adjacent response of that same margin to reach ζ also implies the maximum-response condition used in the proof. The Positional Score summarizes individual query–key curves and supplies a complementary diagnostic of local response. The theorem identifies a spectral constraint when both requirements are imposed on the same margin. The task-level profile itself carries no context-length certificate.

G.1.2 Analytic Finite-Window Moments

For a real angular frequency θ, define

$$K_M(\theta) := \frac{1}{M}\sum_{m=0}^{M-1} e^{im\theta} = \begin{cases} \dfrac{1-e^{iM\theta}}{M(1-e^{i\theta})}, & e^{i\theta} \neq 1, \\ 1, & e^{i\theta} = 1. \end{cases} \tag{175}$$

For $D(m) = c_D + \mathrm{Re}\sum_n d_n e^{im\omega_n}$, let

$$\mu_D = c_D + \mathrm{Re}\sum_n d_n K_M(\omega_n), \tag{176}$$

$$C^-_{n\ell} = K_M(\omega_n - \omega_\ell) - K_M(\omega_n)\overline{K_M(\omega_\ell)}, \tag{177}$$

$$C^+_{n\ell} = K_M(\omega_n + \omega_\ell) - K_M(\omega_n)K_M(\omega_\ell). \tag{178}$$

Expanding the square of the real part gives the exact variance

$$\sigma_D^2 = \frac{1}{2}\,\mathrm{Re}\sum_{n,\ell}\left(d_n\overline{d_\ell}C^-_{n\ell} + d_n d_\ell C^+_{n\ell}\right). \tag{179}$$

Here c_D is the difference of any non-rotary score constants and is zero for fully rotary heads. It contributes to the reference margin $D(0)$ and to μ_D, and leaves the variance unchanged. These expressions require no probability distribution on d. The subsequent Gaussian threshold probability is an approximation to the relative distance law, with the previously established calibration retained.

G.1.3 Relative Ranks and Failure Susceptibility

For a chosen model and head collection, rank a fixed reference set of $N > 1$ task/length settings by decreasing Semantic Score and decreasing Positional Score. Let r_S and r_P be the respective ranks, with rank one best and equal values receiving equal competition ranks. We define the relative semantic and positional failure susceptibilities as

$$w_S = \frac{1}{2} + \frac{r_S - r_P}{2(N-1)}, \qquad w_P = 1 - w_S. \tag{180}$$

A larger w_S indicates a relatively weaker semantic rank, and a larger w_P indicates a relatively weaker positional rank. Equal ranks give $w_S = w_P = 1/2$. These indices describe relative failure susceptibility within the chosen comparison set; their interpretation depends on that set. Raw scores and behavioral success rates remain visible alongside these indices.

G.1.4 Aggregation and Cross-Model Comparison

Semantic reversal probabilities are averaged over valid triplets within each example and head, then over the selected heads, then equally over examples. The Semantic Score is one minus this average probability. For the Positional Score, each pair's adjacent differences are divided by its own coefficient norm before averaging over pairs, heads, and examples. Reference ties satisfy $|D(0)| \leq 64\epsilon_{\mathrm{mach}}\sum_n |d_n|$; their counts and the counts of undefined scores are retained.

For Figure 6, each model ranks all 49 settings by decreasing w_S, assigning average ranks to ties. We compute Spearman correlation as the Pearson correlation between these two vectors of average

ranks, with each setting weighted equally. The resulting correlation is 0.657983, reported as 0.658 in the main text. It includes all 49 settings, including the three adjacent-element retrieval settings highlighted in orange. The plot preserves tied coordinates without jitter. These rankings describe failure susceptibility within the shared reference set.

G.2 Experiment 2: Directed High-Frequency Intervention

G.2.1 Assigning the Optimization Direction

The intervention direction ranks the 49 settings by decreasing w_S, assigning the average rank to ties. Let ρ_S denote this rank. Settings with $\rho_S/49 \leq 1/2$ test reduced high-frequency contributions; the remaining settings test amplification. This places 24 Qwen settings and 23 Llama settings in the reduction group, keeping ties together. For this reference set, reduction corresponds to w_S strictly above the model-specific median: 50.00% for Qwen and 56.25% for Llama. The value $w_S = 1/2$ separately indicates equal semantic and positional ranks.

The reduction group optimizes semantic stability by attenuating high-frequency score contributions or freezing their rotations. The amplification group optimizes positional sensitivity by increasing the high-frequency contribution. The assigned direction remains fixed through both stages of the search.

G.2.2 Intervention Operators and Frequency Mask

For a scalar α, selected high-frequency query and key components are each multiplied by $\sqrt{\alpha}$, which multiplies their score coefficient by α. Low-frequency components and unselected heads retain their original operations. We preserve native query/key normalization and the mapping between query heads and key/value heads. The transformation changes both the coefficient norm and its high-frequency share. A separate no-rotation condition freezes the selected high-frequency RoPE rotations and belongs to the semantic direction.

The high-frequency mask is fixed from prompt length M throughout generation:

$$H = \{n : M\omega_n \geq \gamma\}, \qquad \gamma = \frac{2c_{\text{int}}}{(1-\rho)\epsilon}, \qquad \rho = \theta^{-2/d_{\text{rot}}}, \quad c_{\text{int}} = 0.05, \quad \epsilon = 0.10. \quad (181)$$

Here θ is the RoPE base, d_{rot} is the rotary dimension, and ω_n are the actual runtime frequencies, including native frequency scaling. These intervention runs retain their original empirical parameter $c_{\text{int}} = 0.05$ throughout both search stages. The separate frequency-split audit in Appendix D.2.1 uses $c_{\text{op}} = 0.06$; its later calibration leaves the intervention masks and reported outcomes unchanged.

G.2.3 Two Search Stages

The first stage uses the diagnostic's top-5% head collection: 58 heads for Qwen and 52 for Llama. Semantic-direction settings evaluate no rotation and $\alpha \in \{0, 0.5\}$; positional-direction settings evaluate $\alpha \in \{1.5, 2, 2.5\}$. The unchanged $\alpha = 1$ condition provides the baseline. Candidate and baseline outputs share the original input IDs, greedy decoding, stopping rules, and first-stage output limits. The $\alpha = 1$ implementation was checked against native-model outputs.

The second stage expands the same fixed head ranking to its top 10%: 116 Qwen heads and 103 Llama heads. It selects the 25 Qwen and 23 Llama settings without an observed first-stage improvement in their assigned direction. Semantic-direction settings evaluate no rotation and $\alpha \in \{0.25, 0.5, 0.75\}$; positional-direction settings evaluate $\alpha \in \{1.25, 1.5, 1.75, 2\}$. The diagnostic w_S and task rankings remain fixed at their top-5% values.

G.2.4 Matched Examples, Scoring, and Output Budgets

We preserve the frozen first-stage comparison sets: 2,171 matched examples for Qwen and 2,074 for Llama. Each model has 42 fully covered settings and seven partially covered settings. The second stage completed 8,440 intervention records. Matching its candidates to saved baseline outputs gives 945 Qwen and 1,071 Llama examples; 16 and 78 examples without baseline outputs are excluded. A missing or untested result remains unreported.

Within each stage and setting, every displayed candidate is compared with the original $\alpha = 1$ baseline on the same examples. Accuracy is the fraction of examples receiving an official task score of one.

Appendix H reports the full selected-direction candidate results and matched counts. The stages retain their own comparison sets: Qwen MV (easy) has 94 examples in the first stage and 95 in the second; the other second-stage settings use the same examples as their first-stage comparisons.

Figure 5 intersects the comparison sets across both stages, recomputes their candidates and baseline on this common set, and displays the largest observed gain in the assigned direction. Thus Qwen MV (easy) uses 94 examples for both stages in the figure. The baseline is excluded from the candidate maximum, so settings for which every candidate reduces accuracy retain a negative bar.

The second stage reuses original input IDs and baseline outputs, with shorter generation limits for some tasks. Its reported comparisons retain the original baseline budget. An offline control truncates saved baseline tokens to the shorter limits before rescoring; this changes three Qwen and 28 Llama baseline correctness labels. The original baseline is used in all tables and the main figure.

The search reports the best observed candidate on the examples used to compare intervention strengths. Second-stage settings are selected from first-stage results. These gains describe the available interventions in this exploratory search; choosing an intervention for deployment requires evaluation on held-out examples.

H RESULTS IN THE ASSIGNED OPTIMIZATION DIRECTION

We report every measured candidate in each setting's assigned semantic or positional direction, following the protocol in Appendix G.2. The tables include all 49 settings per model, grouped by direction, with separate results for the top-5% and top-10% head collections.

Observed gains. The first stage improves 24 of 49 Qwen settings and 26 of 49 Llama settings. The second stage adds seven Qwen settings and six Llama settings, giving **31/49 (63.3%) and 32/49 (65.3%)** improved settings, respectively. The largest gains are 20 percentage points for Qwen and 25 for Llama. These are the best observed outcomes from a search over intervention strengths and two head subsets. The selected candidates have not been validated on held-out examples.

Reading the tables. Each table retains its stage's matched examples and original baseline. The count n/N gives matched and planned examples. Candidate headers give α, and NR denotes no rotation. Each entry gives accuracy (%) followed by its change from the paired baseline (percentage points). Dashes indicate settings without a second-stage evaluation, and † marks a second-stage generation limit that differs from the saved baseline. All measured gains, including zero and negative values, are retained. Figure 5 uses the common samples across stages: Qwen MV (easy) uses 94 examples in the figure and 95 in its second-stage table. All other settings use the same paired cohort in the figure and tables. The adjacent-element retrieval settings retain their original 12 examples at each length.

The output-limit control described in Appendix G.2 yields 8/25 and 7/23 improving second-stage settings when the saved baseline outputs are truncated to the second-stage limits before rescoring. The tables retain the original-output comparisons of 7/25 and 6/23.

H.1 QWEN3-8B: SEMANTIC DIRECTION

Table 2: Qwen3-8B: semantic direction on the top-5% heads. Entries give accuracy in percent and the change from the paired baseline in parentheses (percentage points).

Setting	w_S (%)	n/N	Base	NR	0	0.5
QA (basic)	54.2	100/100	78.00	70.00 (-8.00)	76.00 (-2.00)	75.00 (-3.00)
QA (easy)	54.2	100/100	76.00	66.00 (-10.00)	70.00 (-6.00)	77.00 (+1.00)
Multi-hop QA (up to 16k)	53.1	25/25	28.00	24.00 (-4.00)	24.00 (-4.00)	24.00 (-4.00)
Rule deduction (context size 3,000)	51.0	25/25	72.00	88.00 (+16.00)	88.00 (+16.00)	76.00 (+4.00)
Person location (16k)	61.5	25/25	76.00	80.00 (+4.00)	80.00 (+4.00)	68.00 (-8.00)
Previous location (8k)	56.2	25/25	28.00	32.00 (+4.00)	28.00 (+0.00)	32.00 (+4.00)
Previous location (16k)	60.4	25/25	32.00	32.00 (+0.00)	40.00 (+8.00)	36.00 (+4.00)
Directional relations (8k)	70.8	25/25	48.00	44.00 (-4.00)	40.00 (-8.00)	40.00 (-8.00)
Directional relations (16k)	72.9	25/25	48.00	56.00 (+8.00)	56.00 (+8.00)	56.00 (+8.00)
Giving events (8k)	62.5	25/25	72.00	80.00 (+8.00)	84.00 (+12.00)	80.00 (+8.00)
Giving events (16k)	80.2	25/25	60.00	64.00 (+4.00)	64.00 (+4.00)	60.00 (+0.00)
Adjacent-element retrieval (3k)	99.0	12/12	50.00	25.00 (-25.00)	25.00 (-25.00)	33.33 (-16.67)
Adjacent-element retrieval (6k)	95.8	12/12	25.00	8.33 (-16.67)	0.00 (-25.00)	0.00 (-25.00)
Adjacent-element retrieval (13k)	97.9	12/12	16.67	8.33 (-8.33)	8.33 (-8.33)	8.33 (-8.33)
Object location (16k)	74.0	25/25	40.00	40.00 (+0.00)	32.00 (-8.00)	24.00 (-16.00)
Object counting (8k)	68.8	25/25	20.00	24.00 (+4.00)	24.00 (+4.00)	24.00 (+4.00)
Object counting (16k)	77.1	25/25	4.00	20.00 (+16.00)	24.00 (+20.00)	16.00 (+12.00)
Object sets (16k)	56.2	25/25	36.00	36.00 (+0.00)	36.00 (+0.00)	36.00 (+0.00)
Negation (8k)	54.2	25/25	88.00	88.00 (+0.00)	92.00 (+4.00)	92.00 (+4.00)
Negation (16k)	71.9	25/25	68.00	72.00 (+4.00)	76.00 (+8.00)	76.00 (+8.00)
Uncertain knowledge (16k)	62.5	25/25	44.00	48.00 (+4.00)	48.00 (+4.00)	48.00 (+4.00)
One-hop implicit retrieval (16k)	61.5	25/25	12.00	4.00 (-8.00)	8.00 (-4.00)	12.00 (+0.00)
Two-hop implicit retrieval (8k)	65.6	25/25	0.00	0.00 (+0.00)	0.00 (+0.00)	0.00 (+0.00)
Two-hop implicit retrieval (16k)	67.7	25/25	0.00	0.00 (+0.00)	0.00 (+0.00)	0.00 (+0.00)

Table 3: Qwen3-8B: semantic direction on the top-10% heads. Entries give accuracy in percent and the change from the paired baseline in parentheses (percentage points). Dashes mark settings without a second-stage evaluation; † marks a generation limit that differs from the saved baseline.

Setting	n/N	Base	NR	0.25	0.5	0.75
QA (basic)†	100/100	78.00	63.00 (-15.00)	82.00 (+4.00)	80.00 (+2.00)	78.00 (+0.00)
QA (easy)	—	—	—	—	—	—
Multi-hop QA (up to 16k)	25/25	28.00	24.00 (-4.00)	24.00 (-4.00)	24.00 (-4.00)	28.00 (+0.00)
Rule deduction (context size 3,000)	—	—	—	—	—	—
Person location (16k)	—	—	—	—	—	—
Previous location (8k)	—	—	—	—	—	—
Previous location (16k)	—	—	—	—	—	—
Directional relations (8k)	25/25	48.00	44.00 (-4.00)	36.00 (-12.00)	40.00 (-8.00)	52.00 (+4.00)
Directional relations (16k)	—	—	—	—	—	—
Giving events (8k)	—	—	—	—	—	—
Giving events (16k)	—	—	—	—	—	—
Adjacent-element retrieval (3k)	12/12	50.00	41.67 (-8.33)	25.00 (-25.00)	25.00 (-25.00)	33.33 (-16.67)
Adjacent-element retrieval (6k)	12/12	25.00	25.00 (+0.00)	16.67 (-8.33)	16.67 (-8.33)	16.67 (-8.33)
Adjacent-element retrieval (13k)	12/12	16.67	8.33 (-8.33)	8.33 (-8.33)	8.33 (-8.33)	16.67 (+0.00)
Object location (16k)	25/25	40.00	28.00 (-12.00)	28.00 (-12.00)	28.00 (-12.00)	32.00 (-8.00)
Object counting (8k)	—	—	—	—	—	—
Object counting (16k)	—	—	—	—	—	—
Object sets (16k)	25/25	36.00	32.00 (-4.00)	44.00 (+8.00)	44.00 (+8.00)	44.00 (+8.00)
Negation (8k)	—	—	—	—	—	—
Negation (16k)	—	—	—	—	—	—
Uncertain knowledge (16k)	—	—	—	—	—	—
One-hop implicit retrieval (16k)	25/25	12.00	0.00 (-12.00)	12.00 (+0.00)	12.00 (+0.00)	12.00 (+0.00)
Two-hop implicit retrieval (8k)	25/25	0.00	0.00 (+0.00)	0.00 (+0.00)	0.00 (+0.00)	0.00 (+0.00)
Two-hop implicit retrieval (16k)	25/25	0.00	0.00 (+0.00)	0.00 (+0.00)	0.00 (+0.00)	0.00 (+0.00)

H.2 QWEN3-8B: POSITIONAL DIRECTION

Table 4: Qwen3-8B: positional direction on the top-5% heads. Entries give accuracy in percent and the change from the paired baseline in parentheses (percentage points).

Setting	w_S (%)	n/N	Base	1.5	2	2.5
MK (basic)	49.0	100/100	99.00	100.00 (+1.00)	100.00 (+1.00)	100.00 (+1.00)
MK (easy)	16.7	100/100	96.00	97.00 (+1.00)	97.00 (+1.00)	97.00 (+1.00)
MK (medium)	22.9	95/100	72.63	72.63 (+0.00)	70.53 (-2.11)	70.53 (-2.11)
MK (hard)	32.3	95/100	64.21	67.37 (+3.16)	68.42 (+4.21)	70.53 (+6.32)
MV (basic)	46.9	100/100	22.00	15.00 (-7.00)	19.00 (-3.00)	20.00 (-2.00)
MV (easy)	15.6	94/100	43.62	43.62 (+0.00)	42.55 (-1.06)	37.23 (-6.38)
MV (medium)	35.4	94/100	35.11	38.30 (+3.19)	37.23 (+2.13)	36.17 (+1.06)
MV (hard)	22.9	94/100	52.13	53.19 (+1.06)	50.00 (-2.13)	48.94 (-3.19)
QA (medium)	42.7	94/100	69.15	72.34 (+3.19)	71.28 (+2.13)	71.28 (+2.13)
QA (hard)	40.6	94/100	78.72	78.72 (+0.00)	73.40 (-5.32)	76.60 (-2.13)
Paper QA (up to 16k)	12.5	25/25	28.00	20.00 (-8.00)	20.00 (-8.00)	20.00 (-8.00)
Conversation topic retrieval	21.9	150/150	58.00	64.67 (+6.67)	66.00 (+8.00)	64.67 (+6.67)
Text sorting (1k)	12.5	25/25	8.00	8.00 (+0.00)	8.00 (+0.00)	8.00 (+0.00)
Room-property reasoning (context size 500)	16.7	25/25	100.00	100.00 (+0.00)	100.00 (+0.00)	100.00 (+0.00)
Room-property reasoning (context size 3,000)	50.0	25/25	100.00	100.00 (+0.00)	100.00 (+0.00)	100.00 (+0.00)
Relation chaining (context size 500)	20.8	25/25	96.00	92.00 (-4.00)	96.00 (+0.00)	96.00 (+0.00)
Relation chaining (context size 3,000)	41.7	25/25	76.00	72.00 (-4.00)	68.00 (-8.00)	64.00 (-12.00)
Rule deduction (context size 500)	20.8	25/25	84.00	84.00 (+0.00)	72.00 (-12.00)	72.00 (-12.00)
Mixed long reasoning (8k)	45.8	25/25	84.00	80.00 (-4.00)	88.00 (+4.00)	88.00 (+4.00)
Mixed long reasoning (16k)	50.0	25/25	72.00	84.00 (+12.00)	88.00 (+16.00)	84.00 (+12.00)
Person location (8k)	35.4	25/25	92.00	88.00 (-4.00)	96.00 (+4.00)	96.00 (+4.00)
Object location (8k)	37.5	25/25	52.00	48.00 (-4.00)	48.00 (-4.00)	48.00 (-4.00)
Object sets (8k)	34.4	25/25	52.00	36.00 (-16.00)	36.00 (-16.00)	36.00 (-16.00)
Uncertain knowledge (8k)	45.8	25/25	64.00	60.00 (-4.00)	56.00 (-8.00)	56.00 (-8.00)
One-hop implicit retrieval (8k)	50.0	25/25	4.00	12.00 (+8.00)	12.00 (+8.00)	16.00 (+12.00)

Table 5: Qwen3-8B: positional direction on the top-10% heads. Entries give accuracy in percent and the change from the paired baseline in parentheses (percentage points). Dashes mark settings without a second-stage evaluation; † marks a generation limit that differs from the saved baseline.

Setting	n/N	Base	1.25	1.5	1.75	2
MK (basic)	—	—	—	—	—	—
MK (easy)	—	—	—	—	—	—
MK (medium)†	95/100	72.63	70.53 (-2.11)	69.47 (-3.16)	67.37 (-5.26)	67.37 (-5.26)
MK (hard)	—	—	—	—	—	—
MV (basic)†	100/100	22.00	22.00 (+0.00)	24.00 (+2.00)	27.00 (+5.00)	28.00 (+6.00)
MV (easy)	95/100	44.21	40.00 (-4.21)	37.89 (-6.32)	33.68 (-10.53)	28.42 (-15.79)
MV (medium)	—	—	—	—	—	—
MV (hard)	—	—	—	—	—	—
QA (medium)	—	—	—	—	—	—
QA (hard)†	94/100	78.72	79.79 (+1.06)	77.66 (-1.06)	74.47 (-4.26)	72.34 (-6.38)
Paper QA (up to 16k)	25/25	28.00	24.00 (-4.00)	24.00 (-4.00)	24.00 (-4.00)	20.00 (-8.00)
Conversation topic retrieval	—	—	—	—	—	—
Text sorting (1k)	25/25	8.00	8.00 (+0.00)	8.00 (+0.00)	8.00 (+0.00)	8.00 (+0.00)
Room-property reasoning (context size 500)	25/25	100.00	100.00 (+0.00)	100.00 (+0.00)	100.00 (+0.00)	100.00 (+0.00)
Room-property reasoning (context size 3,000)	25/25	100.00	100.00 (+0.00)	100.00 (+0.00)	100.00 (+0.00)	100.00 (+0.00)
Relation chaining (context size 500)	25/25	96.00	92.00 (-4.00)	92.00 (-4.00)	92.00 (-4.00)	92.00 (-4.00)
Relation chaining (context size 3,000)	25/25	76.00	72.00 (-4.00)	72.00 (-4.00)	68.00 (-8.00)	60.00 (-16.00)
Rule deduction (context size 500)	25/25	84.00	84.00 (+0.00)	88.00 (+4.00)	88.00 (+4.00)	84.00 (+0.00)
Mixed long reasoning (8k)	—	—	—	—	—	—
Mixed long reasoning (16k)	—	—	—	—	—	—
Person location (8k)	—	—	—	—	—	—
Object location (8k)	25/25	52.00	56.00 (+4.00)	60.00 (+8.00)	56.00 (+4.00)	56.00 (+4.00)
Object sets (8k)	25/25	52.00	44.00 (-8.00)	40.00 (-12.00)	40.00 (-12.00)	40.00 (-12.00)
Uncertain knowledge (8k)	25/25	64.00	52.00 (-12.00)	44.00 (-20.00)	44.00 (-20.00)	44.00 (-20.00)
One-hop implicit retrieval (8k)	—	—	—	—	—	—

H.3 LLAMA-3.1-8B-INSTRUCT: SEMANTIC DIRECTION

Table 6: Llama-3.1-8B-Instruct: semantic direction on the top-5% heads. Entries give accuracy in percent and the change from the paired baseline in parentheses (percentage points).

Setting	w_S (%)	n/N	Base	NR	0	0.5
QA (easy)	64.6	80/100	80.00	86.25 (+6.25)	87.50 (+7.50)	86.25 (+6.25)
Multi-hop QA (up to 16k)	67.7	25/25	24.00	20.00 (-4.00)	16.00 (-8.00)	16.00 (-8.00)
Person location (16k)	77.1	25/25	76.00	76.00 (+0.00)	84.00 (+8.00)	84.00 (+8.00)
Previous location (8k)	62.5	25/25	28.00	36.00 (+8.00)	28.00 (+0.00)	32.00 (+4.00)
Previous location (16k)	80.2	25/25	20.00	28.00 (+8.00)	24.00 (+4.00)	24.00 (+4.00)
Directional relations (8k)	69.8	25/25	36.00	40.00 (+4.00)	36.00 (+0.00)	40.00 (+4.00)
Directional relations (16k)	70.8	25/25	44.00	56.00 (+12.00)	48.00 (+4.00)	48.00 (+4.00)
Giving events (8k)	57.3	25/25	72.00	72.00 (+0.00)	68.00 (-4.00)	72.00 (+0.00)
Giving events (16k)	58.3	25/25	76.00	64.00 (-12.00)	64.00 (-12.00)	72.00 (-4.00)
Object location (8k)	57.3	25/25	52.00	52.00 (+0.00)	40.00 (-12.00)	44.00 (-8.00)
Object location (16k)	67.7	25/25	36.00	40.00 (+4.00)	32.00 (-4.00)	32.00 (-4.00)
Object counting (8k)	65.6	25/25	4.00	12.00 (+8.00)	8.00 (+4.00)	12.00 (+8.00)
Object counting (16k)	80.2	25/25	8.00	12.00 (+4.00)	16.00 (+8.00)	12.00 (+4.00)
Object sets (8k)	57.3	25/25	44.00	44.00 (+0.00)	52.00 (+8.00)	44.00 (+0.00)
Object sets (16k)	67.7	25/25	40.00	48.00 (+8.00)	48.00 (+8.00)	48.00 (+8.00)
Negation (8k)	58.3	25/25	88.00	88.00 (+0.00)	84.00 (-4.00)	84.00 (-4.00)
Negation (16k)	76.0	25/25	72.00	64.00 (-8.00)	72.00 (+0.00)	72.00 (+0.00)
Uncertain knowledge (8k)	71.9	25/25	76.00	80.00 (+4.00)	72.00 (-4.00)	72.00 (-4.00)
Uncertain knowledge (16k)	80.2	25/25	64.00	68.00 (+4.00)	56.00 (-8.00)	68.00 (+4.00)
One-hop implicit retrieval (8k)	70.8	25/25	36.00	40.00 (+4.00)	24.00 (-12.00)	40.00 (+4.00)
One-hop implicit retrieval (16k)	83.3	25/25	24.00	28.00 (+4.00)	36.00 (+12.00)	28.00 (+4.00)
Two-hop implicit retrieval (8k)	72.9	25/25	4.00	0.00 (-4.00)	8.00 (+4.00)	8.00 (+4.00)
Two-hop implicit retrieval (16k)	70.8	25/25	0.00	4.00 (+4.00)	0.00 (+0.00)	4.00 (+4.00)

Table 7: Llama-3.1-8B-Instruct: semantic direction on the top-10% heads. Entries give accuracy in percent and the change from the paired baseline in parentheses (percentage points). Dashes mark settings without a second-stage evaluation; † marks a generation limit that differs from the saved baseline.

Setting	n/N	Base	NR	0.25	0.5	0.75
QA (easy)	—	—	—	—	—	—
Multi-hop QA (up to 16k)	25/25	24.00	24.00 (+0.00)	8.00 (-16.00)	8.00 (-16.00)	16.00 (-8.00)
Person location (16k)	—	—	—	—	—	—
Previous location (8k)	—	—	—	—	—	—
Previous location (16k)	—	—	—	—	—	—
Directional relations (8k)	—	—	—	—	—	—
Directional relations (16k)	—	—	—	—	—	—
Giving events (8k)	25/25	72.00	72.00 (+0.00)	68.00 (-4.00)	72.00 (+0.00)	72.00 (+0.00)
Giving events (16k)	25/25	76.00	68.00 (-8.00)	64.00 (-12.00)	68.00 (-8.00)	72.00 (-4.00)
Object location (8k)	25/25	52.00	20.00 (-32.00)	48.00 (-4.00)	40.00 (-12.00)	44.00 (-8.00)
Object location (16k)	—	—	—	—	—	—
Object counting (8k)	—	—	—	—	—	—
Object counting (16k)	—	—	—	—	—	—
Object sets (8k)	—	—	—	—	—	—
Object sets (16k)	—	—	—	—	—	—
Negation (8k)	25/25	88.00	80.00 (-8.00)	76.00 (-12.00)	84.00 (-4.00)	84.00 (-4.00)
Negation (16k)	25/25	72.00	52.00 (-20.00)	60.00 (-12.00)	64.00 (-8.00)	68.00 (-4.00)
Uncertain knowledge (8k)	—	—	—	—	—	—
Uncertain knowledge (16k)	—	—	—	—	—	—
One-hop implicit retrieval (8k)	—	—	—	—	—	—
One-hop implicit retrieval (16k)	—	—	—	—	—	—
Two-hop implicit retrieval (8k)	—	—	—	—	—	—
Two-hop implicit retrieval (16k)	—	—	—	—	—	—

H.4 LLAMA-3.1-8B-INSTRUCT: POSITIONAL DIRECTION

Table 8: Llama-3.1-8B-Instruct: positional direction on the top-5% heads. Entries give accuracy in percent and the change from the paired baseline in parentheses (percentage points).

Setting	w_S (%)	n/N	Base	1.5	2	2.5
MK (basic)	55.2	81/100	98.77	100.00 (+1.23)	97.53 (-1.23)	27.16 (-71.60)
MK (easy)	11.5	100/100	83.00	80.00 (-3.00)	5.00 (-78.00)	0.00 (-83.00)
MK (medium)	8.3	100/100	61.00	61.00 (+0.00)	24.00 (-37.00)	1.00 (-60.00)
MK (hard)	20.8	81/100	54.32	48.15 (-6.17)	1.23 (-53.09)	0.00 (-54.32)
MV (basic)	54.2	100/100	34.00	20.00 (-14.00)	7.00 (-27.00)	0.00 (-34.00)
MV (easy)	14.6	81/100	27.16	19.75 (-7.41)	0.00 (-27.16)	0.00 (-27.16)
MV (medium)	11.5	80/100	32.50	45.00 (+12.50)	3.75 (-28.75)	0.00 (-32.50)
MV (hard)	12.5	100/100	24.00	23.00 (-1.00)	3.00 (-21.00)	0.00 (-24.00)
QA (basic)	54.2	80/100	97.50	97.50 (+0.00)	45.00 (-52.50)	0.00 (-97.50)
QA (medium)	56.2	80/100	67.50	66.25 (-1.25)	36.25 (-31.25)	2.50 (-65.00)
QA (hard)	56.2	100/100	71.00	72.00 (+1.00)	35.00 (-36.00)	12.00 (-59.00)
Paper QA (up to 16k)	31.2	25/25	20.00	20.00 (+0.00)	20.00 (+0.00)	0.00 (-20.00)
Conversation topic retrieval	17.7	150/150	74.00	80.00 (+6.00)	35.33 (-38.67)	0.00 (-74.00)
Text sorting (1k)	10.4	25/25	0.00	0.00 (+0.00)	0.00 (+0.00)	0.00 (+0.00)
Room-property reasoning (context size 500)	19.8	25/25	64.00	60.00 (-4.00)	44.00 (-20.00)	56.00 (-8.00)
Room-property reasoning (context size 3,000)	56.2	25/25	44.00	44.00 (+0.00)	60.00 (+16.00)	48.00 (+4.00)
Relation chaining (context size 500)	21.9	25/25	52.00	52.00 (+0.00)	56.00 (+4.00)	56.00 (+4.00)
Relation chaining (context size 3,000)	45.8	25/25	56.00	52.00 (-4.00)	48.00 (-8.00)	48.00 (-8.00)
Rule deduction (context size 500)	15.6	25/25	60.00	52.00 (-8.00)	48.00 (-12.00)	48.00 (-12.00)
Rule deduction (context size 3,000)	42.7	25/25	48.00	48.00 (+0.00)	56.00 (+8.00)	40.00 (-8.00)
Mixed long reasoning (8k)	47.9	25/25	76.00	56.00 (-20.00)	56.00 (-20.00)	8.00 (-68.00)
Mixed long reasoning (16k)	39.6	25/25	64.00	68.00 (+4.00)	48.00 (-16.00)	4.00 (-60.00)
Person location (8k)	54.2	25/25	96.00	92.00 (-4.00)	44.00 (-52.00)	0.00 (-96.00)
Adjacent-element retrieval (3k)	26.0	12/12	0.00	0.00 (+0.00)	0.00 (+0.00)	0.00 (+0.00)
Adjacent-element retrieval (6k)	37.5	12/12	0.00	0.00 (+0.00)	8.33 (+8.33)	0.00 (+0.00)
Adjacent-element retrieval (13k)	39.6	12/12	16.67	8.33 (-8.33)	8.33 (-8.33)	0.00 (-16.67)

Table 9: Llama-3.1-8B-Instruct: positional direction on the top-10% heads. Entries give accuracy in percent and the change from the paired baseline in parentheses (percentage points). Dashes mark settings without a second-stage evaluation; † marks a generation limit that differs from the saved baseline.

Setting	n/N	Base	1.25	1.5	1.75	2
MK (basic)	—	—	—	—	—	—
MK (easy)†	100/100	83.00	79.00 (-4.00)	71.00 (-12.00)	68.00 (-15.00)	61.00 (-22.00)
MK (medium)†	100/100	61.00	60.00 (-1.00)	57.00 (-4.00)	47.00 (-14.00)	31.00 (-30.00)
MK (hard)†	81/100	54.32	44.44 (-9.88)	40.74 (-13.58)	23.46 (-30.86)	11.11 (-43.21)
MV (basic)†	100/100	34.00	19.00 (-15.00)	5.00 (-29.00)	1.00 (-33.00)	6.00 (-28.00)
MV (easy)	81/100	27.16	29.63 (+2.47)	30.86 (+3.70)	12.35 (-14.81)	0.00 (-27.16)
MV (medium)	—	—	—	—	—	—
MV (hard)†	100/100	24.00	19.00 (-5.00)	32.00 (+8.00)	30.00 (+6.00)	18.00 (-6.00)
QA (basic)†	80/100	97.50	90.00 (-7.50)	86.25 (-11.25)	66.25 (-31.25)	13.75 (-83.75)
QA (medium)†	80/100	67.50	41.25 (-26.25)	28.75 (-38.75)	38.75 (-28.75)	23.75 (-43.75)
QA (hard)	—	—	—	—	—	—
Paper QA (up to 16k)	25/25	20.00	20.00 (+0.00)	20.00 (+0.00)	20.00 (+0.00)	24.00 (+4.00)
Conversation topic retrieval	—	—	—	—	—	—
Text sorting (1k)	25/25	0.00	0.00 (+0.00)	8.00 (+8.00)	8.00 (+8.00)	16.00 (+16.00)
Room-property reasoning (context size 500)	25/25	64.00	52.00 (-12.00)	48.00 (-16.00)	52.00 (-12.00)	48.00 (-16.00)
Room-property reasoning (context size 3,000)	—	—	—	—	—	—
Relation chaining (context size 500)	—	—	—	—	—	—
Relation chaining (context size 3,000)	25/25	56.00	52.00 (-4.00)	52.00 (-4.00)	52.00 (-4.00)	60.00 (+4.00)
Rule deduction (context size 500)	25/25	60.00	52.00 (-8.00)	52.00 (-8.00)	56.00 (-4.00)	52.00 (-8.00)
Rule deduction (context size 3,000)	—	—	—	—	—	—
Mixed long reasoning (8k)	25/25	76.00	56.00 (-20.00)	48.00 (-28.00)	40.00 (-36.00)	36.00 (-40.00)
Mixed long reasoning (16k)	—	—	—	—	—	—
Person location (8k)	25/25	96.00	92.00 (-4.00)	76.00 (-20.00)	84.00 (-12.00)	32.00 (-64.00)
Adjacent-element retrieval (3k)	12/12	0.00	0.00 (+0.00)	8.33 (+8.33)	25.00 (+25.00)	8.33 (+8.33)
Adjacent-element retrieval (6k)	—	—	—	—	—	—
Adjacent-element retrieval (13k)	12/12	16.67	16.67 (+0.00)	16.67 (+0.00)	0.00 (-16.67)	0.00 (-16.67)

I Additional Related Work

I.1 Representation Assumptions in RoPE Theory

Existing RoPE theories study both random representations and fixed content vectors. Barbero et al. (2025) use independently sampled Gaussian queries and keys to show that expected scores need not decay with distance; their paper also provides deterministic constructions and a single-frequency semantic-instability result. Xu et al. (2024) derive base-dependent context bounds using i.i.d. query and key coordinates with a common variance and a similar-key model $k^* = q + \epsilon$ with zero-mean noise. Du et al. (2026) hold query and key vectors fixed and sample relative positions. Their normal approximation assumes that no frequency amplitude dominates, and their context-dependent failure estimates use additional amplitude regularity. For their theoretical token-inversion estimate, they compare a key with a hypothetical phase-aligned reference sharing its frequency amplitudes. Our analysis uses the coefficients of realized query and key activations, without prescribing a distribution over these activations. The finite-window moment and adjacent-response bounds allow nonuniform frequency amplitudes. Our diagnostic estimates reversal for sampled key pairs using their full score-margin coefficients. The reversal guarantees retain an explicit Gaussian approximation condition on the score distribution over positions. This formulation supports diagnostics computed directly from recorded model activations.

I.2 RoPE Failure Mechanisms

Existing analyses identify failures in both the positional and semantic functions of rotary attention. Barbero et al. (2025) show how rotating semantic channels can lose their selectivity over long contexts, with a formal result for a single frequency, and motivate p-RoPE. Du et al. (2026) organize score-level failures into four types. *Position inversion* occurs when the same query–key pair scores higher at a distant position than at a nearby one, reversing the locality preference. *Position aliasing* occurs when that pair receives numerically equal scores at two different relative distances. For two distinct keys evaluated at the same relative distance, *token inversion* reverses the ordering established at zero relative distance, while *token aliasing* makes their scores numerically equal. The aliasing events explicitly account for finite-precision computation. These definitions separate loss of positional discrimination from loss of token discrimination, and distinguish a reversed preference from a tie.

In vision–language models, Qi et al. (2026) probe positional sensitivity by adding a one-step RoPE rotation to visual keys while holding their extracted content vectors and the query fixed. They report changes in attention weights and in group-averaged score sensitivity. This study highlights local positional resolution as a practical concern. In §2, our semantic reversal follows the token-inversion event of Du et al. (2026). We define positional insensitivity through the absolute score gap between adjacent positions, normalized by each query–key pair's own coefficient norm. This criterion measures local response strength independently of its direction; a preference for nearer positions and equal scores at separated positions are distinct properties.

I.3 Context Extension

Other work explains the mechanisms behind context-extension failures. HoPE connects extrapolation failures to learned frequency components whose attention-score patterns change beyond the training window (Chen et al., 2025). Resonance RoPE identifies a feature gap at unseen integer positions, including in frequency components whose value ranges are covered during training (Wang et al., 2024b). Frayed RoPE relates long-context degradation to dispersion and overlap of query–key clusters, which weaken the ability of attention-sink tokens to absorb attention (Wertheimer et al., 2026). These accounts connect frequency behavior and representation geometry to specific failures, complementing the event-based classification in Appendix I.2.

Position Interpolation and YaRN adjust RoPE frequencies to extend context (Chen et al., 2023b; Peng et al., 2024). For position interpolation, Wu et al. (2026) show that its practical effect depends on the task: frequency scaling can help dependencies that stretch with input length while impairing retrieval at fixed relative distances and precise local alignment. YaRN also identifies the difficulty of distinguishing nearby positions after uniform frequency scaling (Peng et al., 2024).

I.4 Internal Representations and Failure Diagnosis

Dong et al. (2024) decompose hidden states using position-wise means across sequences, connect positional-vector extrapolation to failures, and develop interventions for NoPE models. Other work diagnoses positional attention bias (Hsieh et al., 2024b), retrieval-head behavior (Wu et al., 2025a), and context-extension mechanisms through head ablations and activation patching (Zhao et al., 2025). StreamingLLM links rolling-cache degradation to attention sinks (Xiao et al., 2024), and contrastive attribution traces output errors through internal states (Tan et al., 2026). Our local score measurements complement these analyses; relating a measured vulnerability to an output error requires evidence about its downstream effect.

I.5 Behavioral Evaluation and Attention Competition

LongBench, RULER, and HELMET measure performance across long-context tasks (Bai et al., 2024; Hsieh et al., 2024a; Yen et al., 2025). Lost in the Middle and NoLiMa expose sensitivity to information placement and retrieval beyond literal matching (Liu et al., 2024a; Modarressi et al., 2025). Longer inputs motivate analyses of distractor dilution (Bansal et al., 2026) and critical attention scaling under simplified token geometries (Chen et al., 2026). Parallel context encoding has also been linked to elevated attention entropy (Zhang et al., 2025). Our pre-softmax diagnostics add two internal measurements alongside task performance; softmax competition and subsequent layers determine how these properties influence the final prediction.

I.6 Long-Context Failure Modes

Existing benchmarks and studies (Hsieh et al., 2024a; Bai et al., 2024; Yen et al., 2025; Kuratov et al., 2024; Modarressi et al., 2025; Goldman et al., 2024; Vodrahalli et al., 2024; Levy et al., 2024) identify long-context failures related to retrieval and reasoning, and that these failure modes can be coupled (Du et al., 2025). Our diagnostics feature two internal measurements and position the model on a preference spectrum with preset tasks anchored as combinations of both modes.