

Rotary Position Embedding (RoPE) & Context Extrapolation

Mathematical Mechanics of Frequency Drift, NTK-Aware Scaling, and YaRN in 1M+ Token LLMs

Published by *Lesty Tech Machine Learning Research Lab* | Date: October 19, 2026 | Reference: LESTY-ENG-2026-B5

1. Mathematical Foundations of Rotary Position Embedding

Rotary Position Embedding (RoPE) encodes relative positional token dependencies by rotating query and key representations in complex two-dimensional sub-spaces. The rotation angle for position index m at hidden channel dimension i is defined by $\theta_i = b^{-2(i-1)/d}$, where $b=10000$ and d is the head dimension. Because inner products depend strictly on relative distance $(m - n)$, RoPE provides clean theoretical decay of attention scores across sequences within the pre-trained context limit.

2. The Frequency Drift Dilemma in Context Extrapolation

Mathematical Failure Mode: When context is naively extended from 8,192 to 131,072 tokens using linear Position Interpolation (PI), all frequencies are scaled by a constant factor $s=16$. This causes high-frequency components (which govern adjacent word grammar and local token semantics) to blur together, destroying the model's ability to perform precise syntax parsing. Conversely, naive extrapolation without scaling causes out-of-distribution phase rotations on low-frequency components, causing attention logits to explode.

3. The YaRN (Yet another RoPE extension) Solution

YaRN divides the RoPE frequency spectrum into three distinct non-linear wavelength bands: (1) high-frequency wavelengths shorter than pretraining context length are left unscaled (pure extrapolation); (2) low-frequency wavelengths significantly longer than context length are fully interpolated; and (3) intermediate wavelengths undergo smooth ramp-function blending. Combined with a temperature-scaling factor $\sqrt{t}$ to counteract attention score flattening, YaRN scales LLM attention to 1M+ tokens without perplexity degradation.

ARCHITECTURAL BENCHMARK & PARAMETER TELEMETRY

Scaling Strategy	Linear Interpolation (PI)	NTK-Aware RoPE	YaRN (Non-Linear + Temp)
High-Frequency Scaling	Uniformly diluted (1/s)	Base adjusted: $b' = b * s^{(d/(d-2))}$	Zero scaling (Preserves local syntax)
Low-Frequency Scaling	Uniformly diluted (1/s)	Moderate compression	Full interpolation (Wavelength match)
Attention Logit Entropy	+42% Flattening (Perplexity loss)	+18% Minor dispersion	Normalized via temperature correction
Passkey Retrieval (128K)	64.2% Accuracy	88.7% Accuracy	99.8% Accuracy

Fine-Tuning Steps Required	~2,000 steps	~1,200 steps	< 400 steps
----------------------------	--------------	--------------	-------------

© 2026 Lesty.tech Semiconductor & Systems Architecture Group. Released under Creative Commons Technical Attribution 4.0 International. For peer review inquiries and technical data verification, visit <https://lesty.tech>