Rotary Position Embeddings (RoPE)

Comprehensive Demo and Analysis

RoPE encodes relative position directly into attention dot products through rotation. Each dimension pair defines a 2D subspace with position-dependent rotation.

Key property: <RoPE(q,m), RoPE(k,n)> depends only on (m-n), making RoPE a true relative position encoding.

Used by: Llama 1/2/3, Mistral, Qwen, Gemma, DeepSeek, and virtually all modern open-weight LLMs.

Random seed: 42
Number of visualizations: 6
Examples: 6

Generated by demo.py

Summary of Findings

1. Rotation Visualization: Each dimension pair (2i, 2i+1) traces a circle in its 2D subspace. Pair 0 rotates fast (theta_0 = 1.0), higher pairs rotate progressively slower. Norm is preserved because rotations are orthogonal transformations.

2. Relative Position Property (CENTERPIECE): The dot product <RoPE(q,m), RoPE(k,n)> = q^T R(n-m) k depends ONLY on (m-n). Proof: R(m)^T R(n) = R(-m)R(n) = R(n-m) by angle addition. Verified empirically: max variation < 1e-12 across all positions.

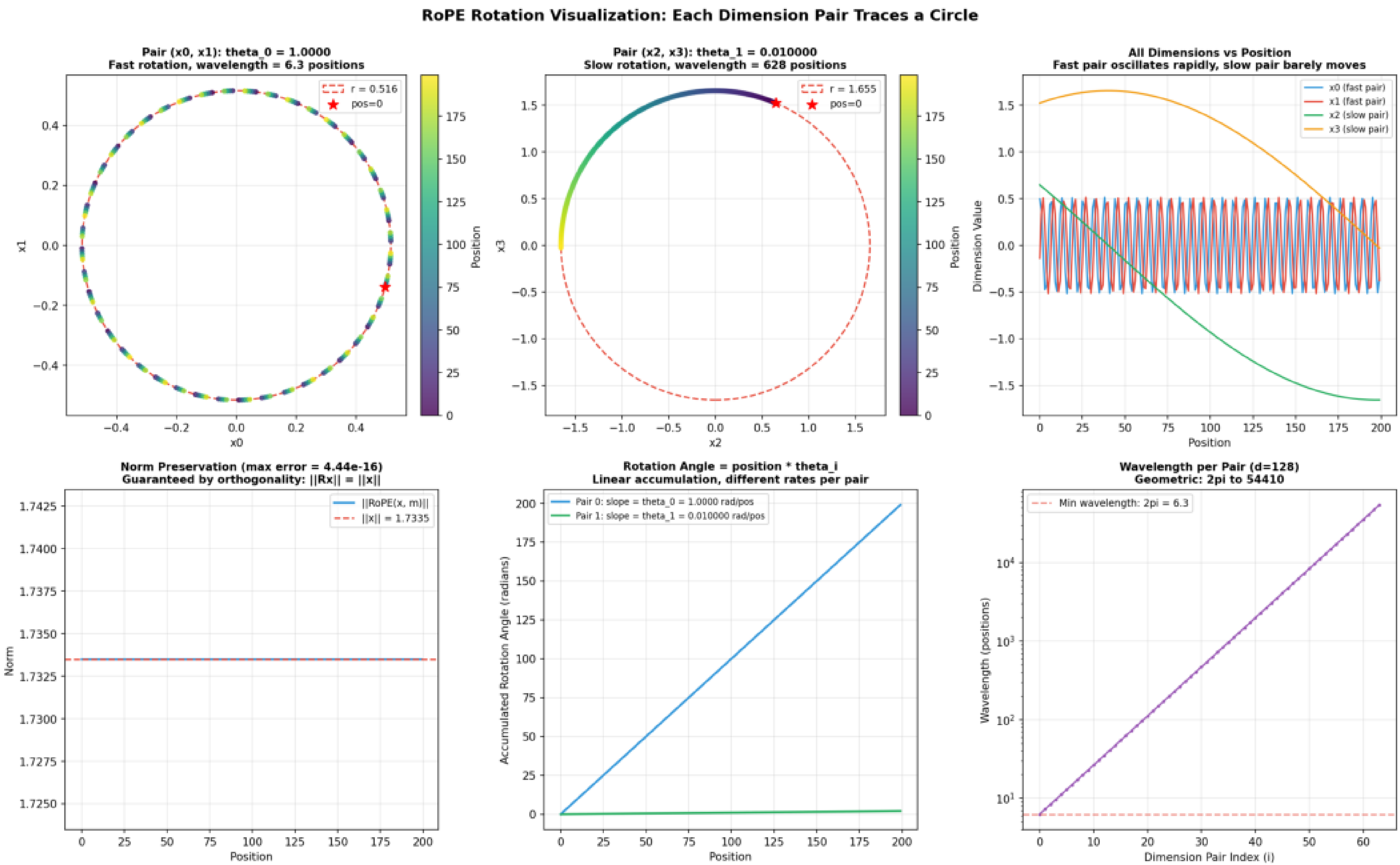
3. Norm & Orthogonality: ||RoPE(x,m)|| = ||x|| guaranteed by R^T R = I. det(R) = 1 (proper rotation). R(m)R(n) = R(m+n) (composition). R(m)R(-m) = I (inverse). All analytically exact from cos^2+sin^2=1.

4. RoPE vs Sinusoidal PE: Both use theta_i = 10000^(-2i/d). Sinusoidal is ADDITIVE (position added to content), creating cross-terms that depend on absolute position. RoPE is MULTIPLICATIVE (q' = R(m)q), giving a pure relative position encoding. This is the key advantage.

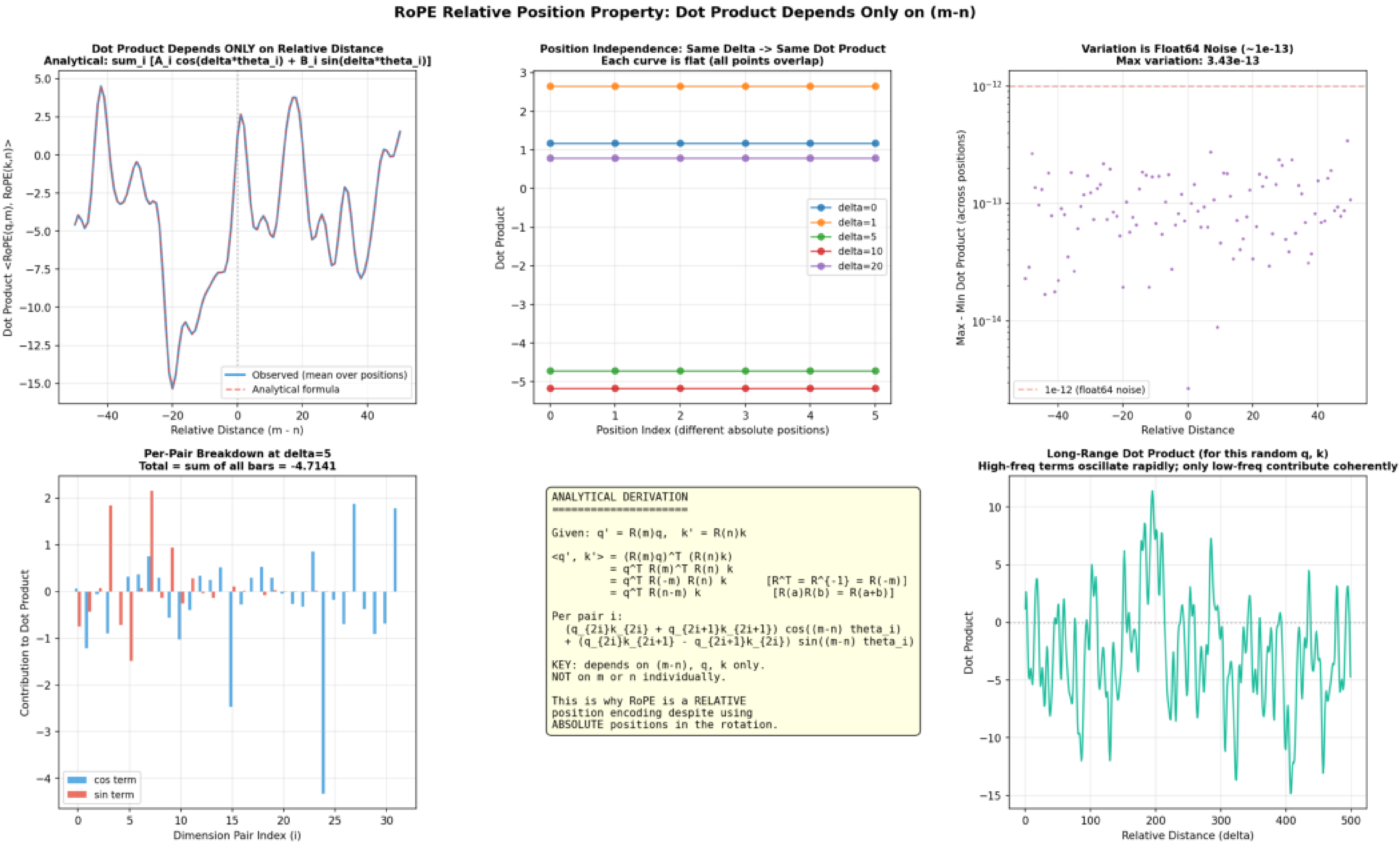
5. Attention Impact: With identical token embeddings, RoPE creates position-dependent attention patterns (weights vary by relative position). Without RoPE, attention is uniform. Patterns are shift-invariant (shifting all positions preserves relative distances).

6. Context Extension: Larger theta_base stretches wavelengths, extending effective context. Llama 3 uses theta=500K vs standard 10K. NTK-aware scaling preserves high-freq (local) while stretching low-freq (long-range) -- can extend context without retraining.

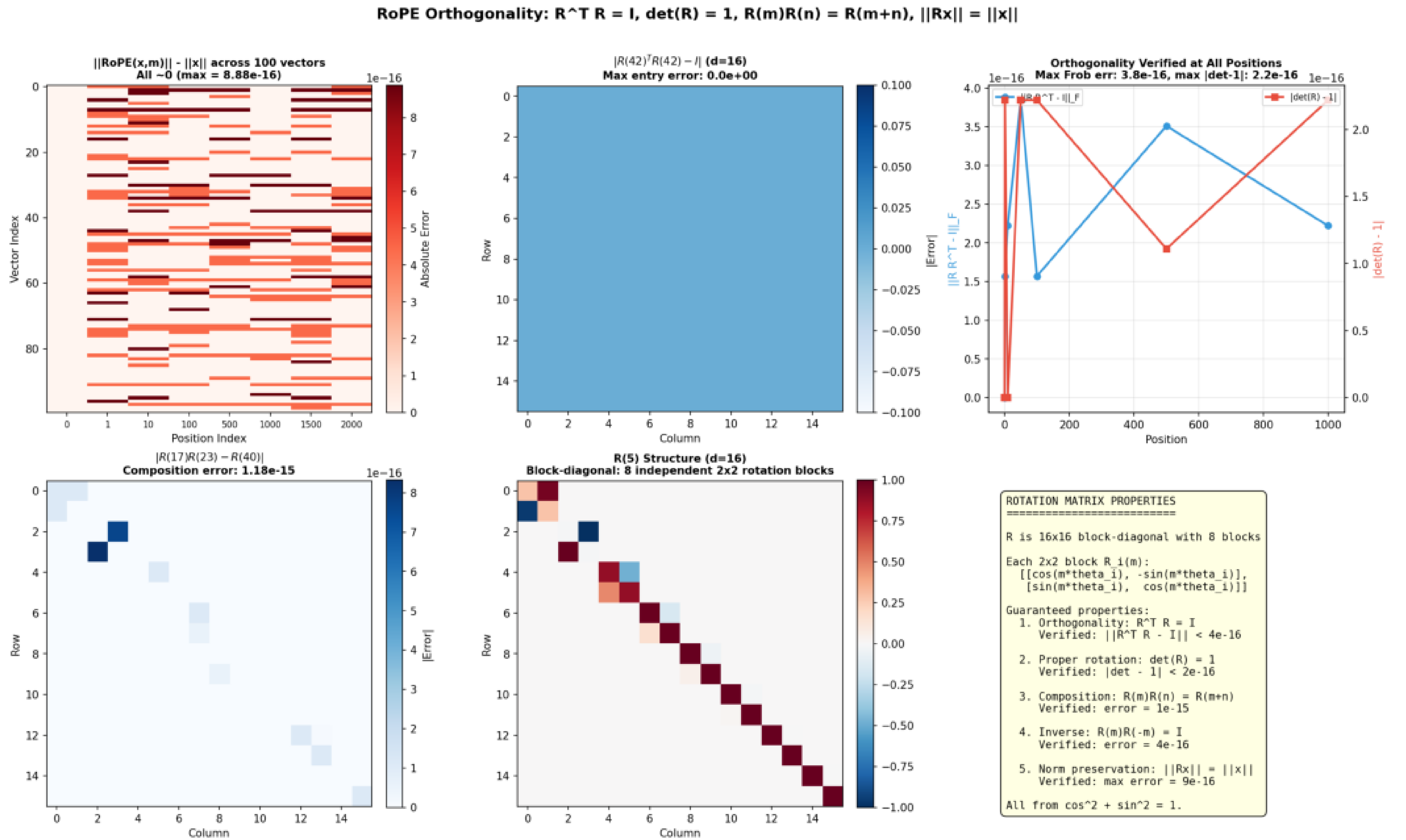
Example 1: Rotation Visualization -- Circles in 2D Subspaces



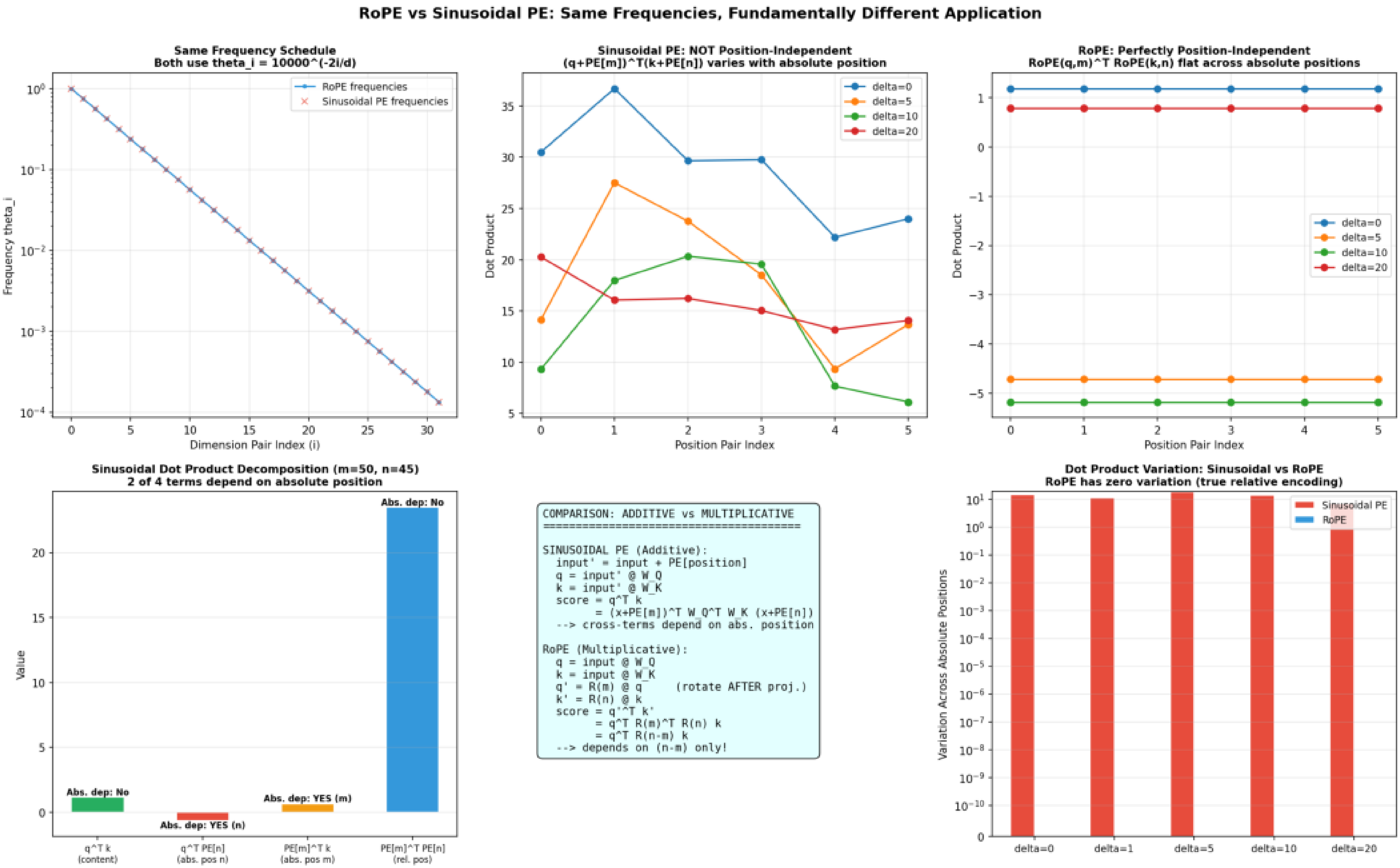
Example 2: Relative Position Property -- The Key Theorem



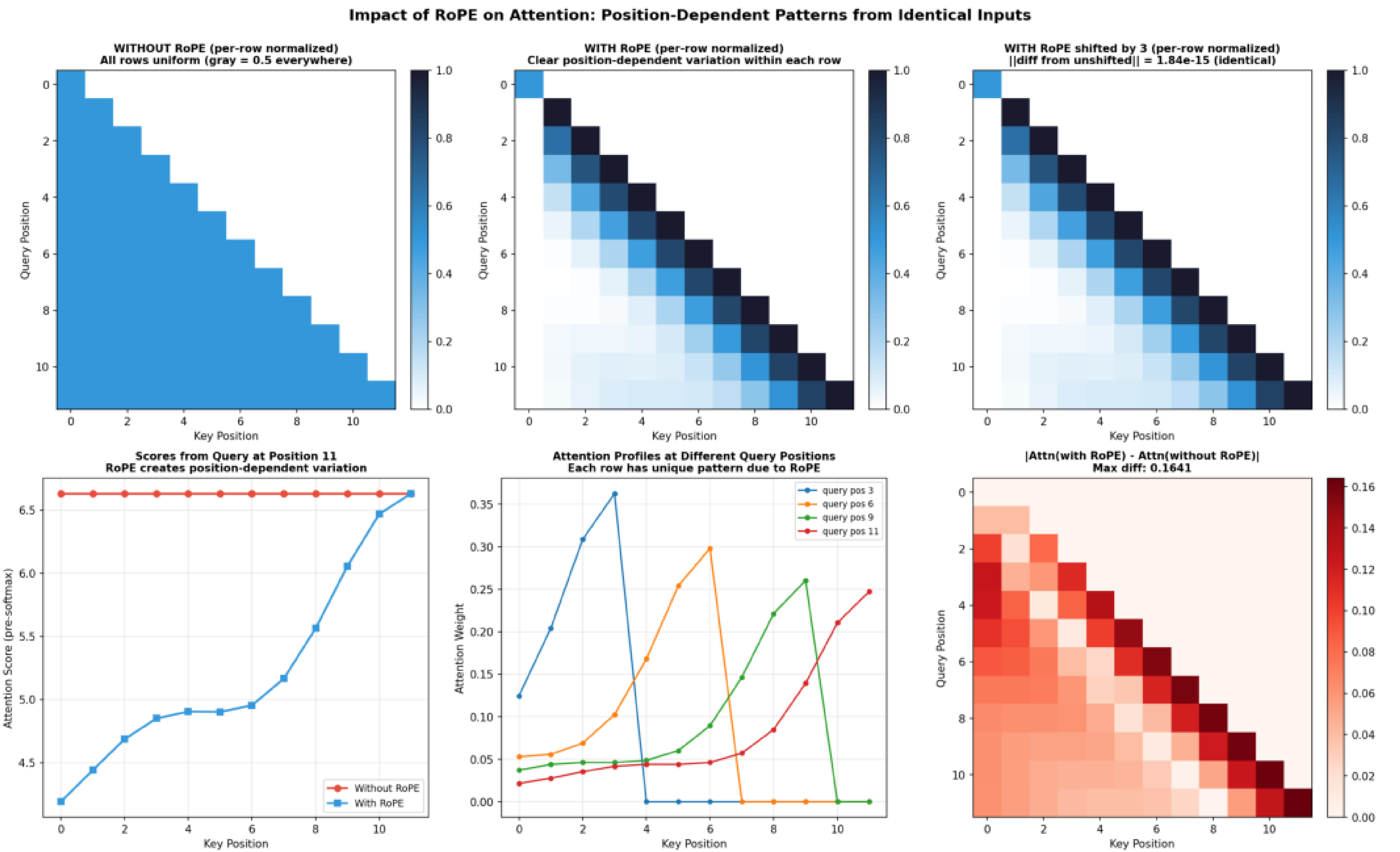
Example 3: Norm Preservation & Orthogonality



Example 4: RoPE vs Sinusoidal PE -- Additive vs Multiplicative



Example 5: Impact on Attention Patterns



Example 6: Context Extension via Theta Scaling