Transformer Block

Pre-Norm Decoder Block: RMSNorm + GQA + RoPE + SwiGLU

The fundamental repeated unit of every modern LLM.
Wires together RMSNorm, grouped-query attention with RoPE,
and a SwiGLU FFN into the pre-norm architecture used by
Llama, Mistral, and all modern open-weight models.

This demo covers:

1. Full forward pass walkthrough with shape tracing
2. Parameter distribution for Llama 2/3 and Mistral configs
3. SwiGLU gating mechanism visualization
4. Residual connection gradient highway analysis
5. FLOPs breakdown with analytical crossover points
6. Causal masking and RoPE position sensitivity

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

Generated by demo.py

Summary of Findings

1. Forward Pass: Output shape matches input (B, L, d_model). Pre-norm architecture normalizes inputs to sublayers (RMS ~ 1.0) while the residual stream remains unnormalized and grows with depth.

2. Parameter Distribution: FFN (SwiGLU) dominates per-block parameters: ~67% with MHA (Llama 2 7B), ~80%+ with GQA (70B, Llama 3, Mistral). GQA reduces attention params (Llama 2 70B saves ~44% vs MHA by using 8 KV heads for 64 Q heads). Norms are negligible (<0.01%).

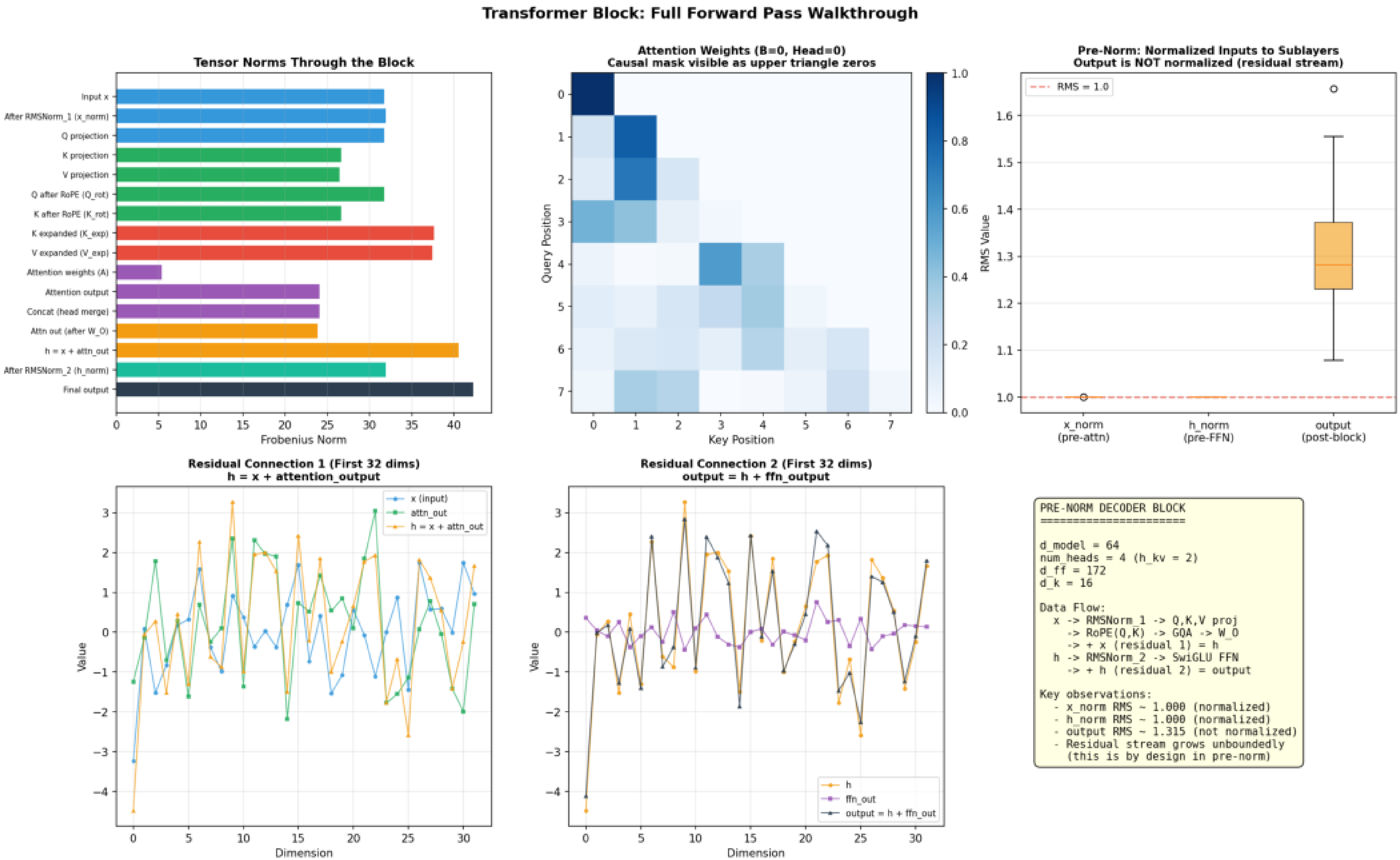
3. SwiGLU Gating: The gate signal SiLU(x @ W_gate) selectively suppresses features. SiLU is smooth (no dead neurons unlike ReLU). The gating mechanism enables learned feature selection. 3 matrices vs 2, but d_ff = 8/3 * d compensates for the extra parameters.

4. Residual Connections: The 'gradient highway' ensures gradients never vanish regardless of depth. Pre-norm gives d(output)/d(x) = I + ... The identity term persists through all layers. With 1/sqrt(N) weight scaling, norms grow moderately and gradients remain stable.

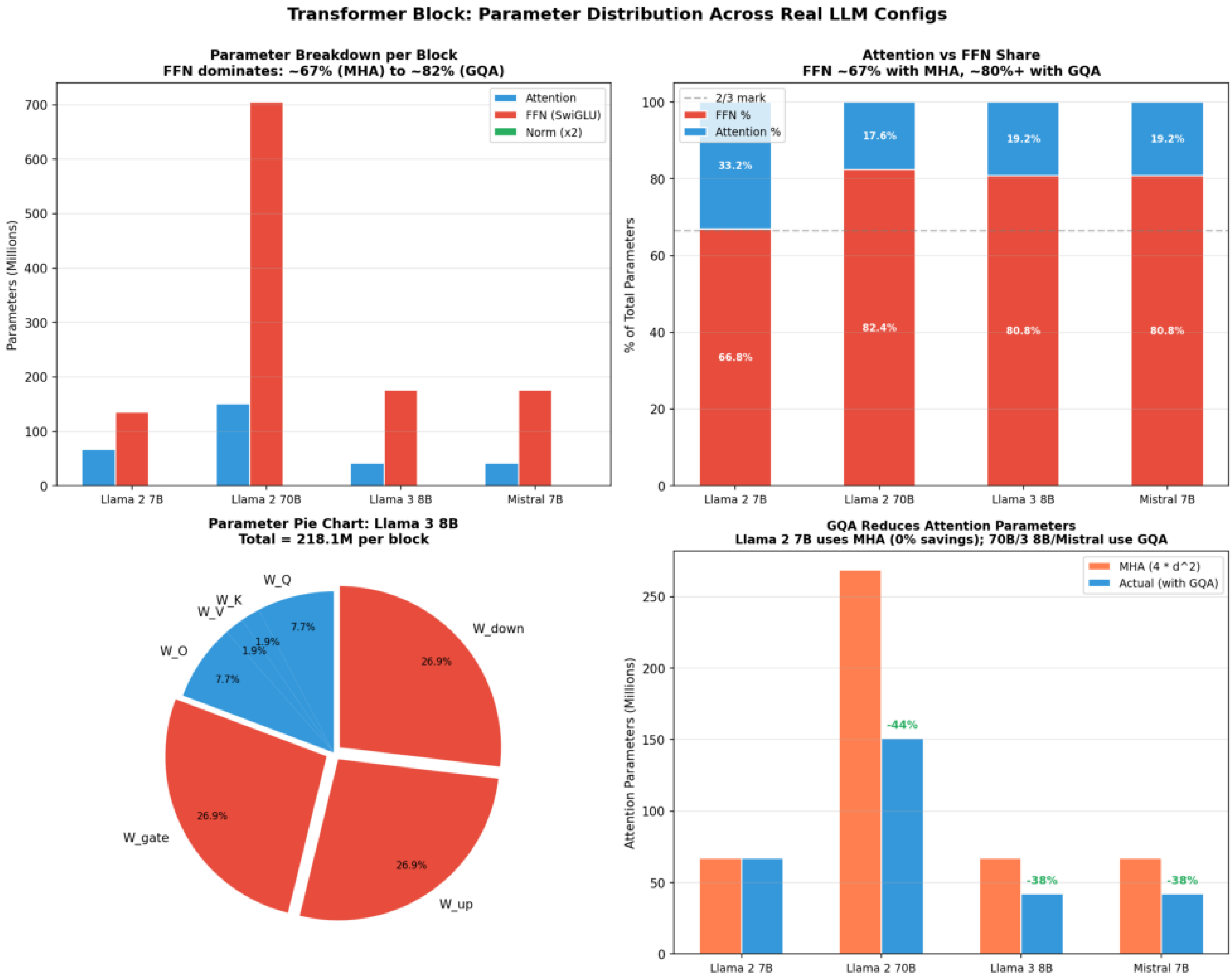
5. FLOPs: Attention core is O(L^2) while FFN and projections are O(L). For Llama 3 8B, attention core surpasses FFN at L ~ 21,296 tokens. Formula: L_cross = 6*d*d_ff / ((4*d_k+5)*h). At short sequences, FFN dominates; at long sequences, attention core dominates.

6. Causal Masking: Position i's output depends only on positions 0..i. Verified by showing that changing future tokens has zero effect on past outputs. RoPE makes identical tokens position-aware through rotation, even without any trained weights.

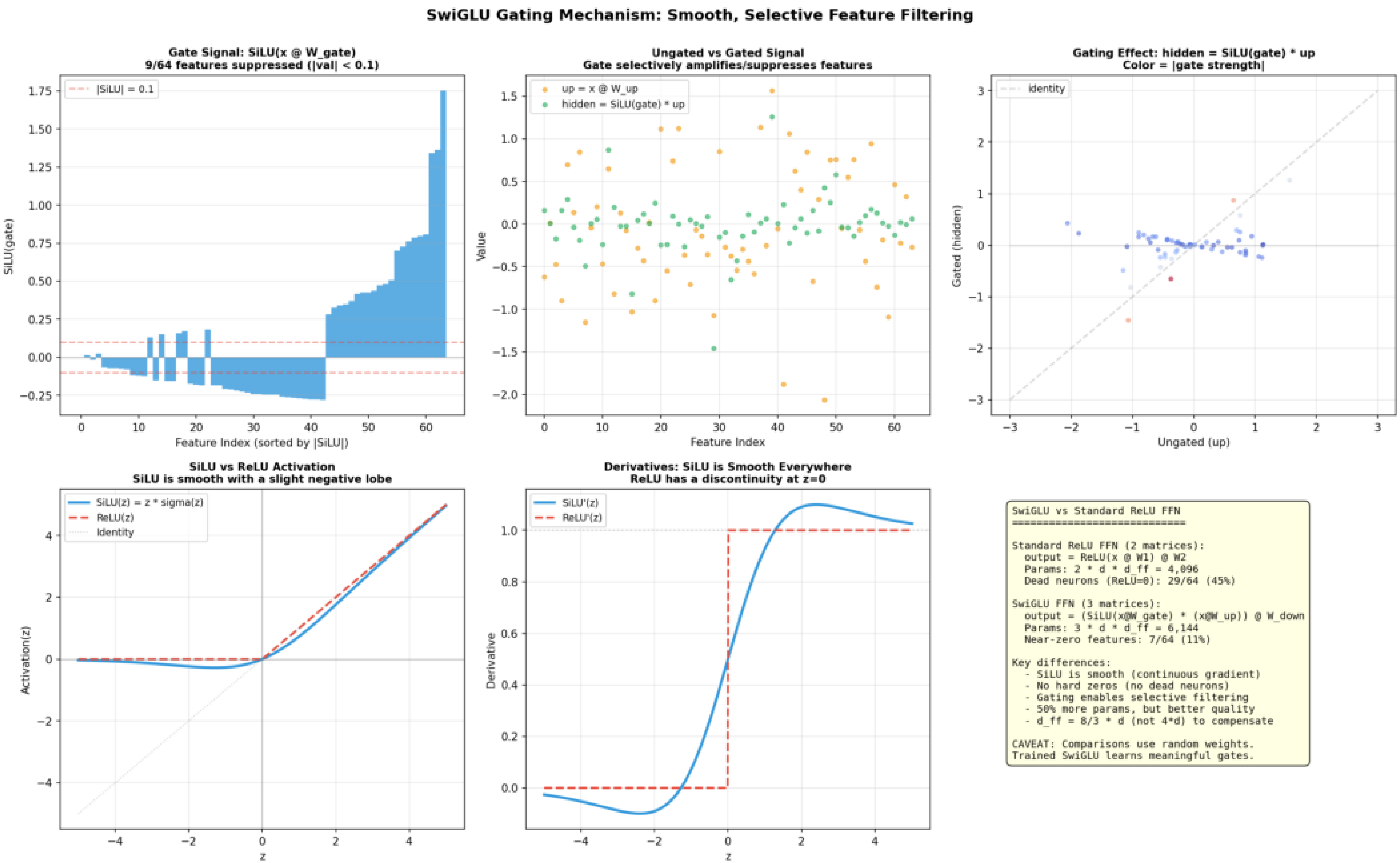
Example 1: Full Forward Pass Walkthrough



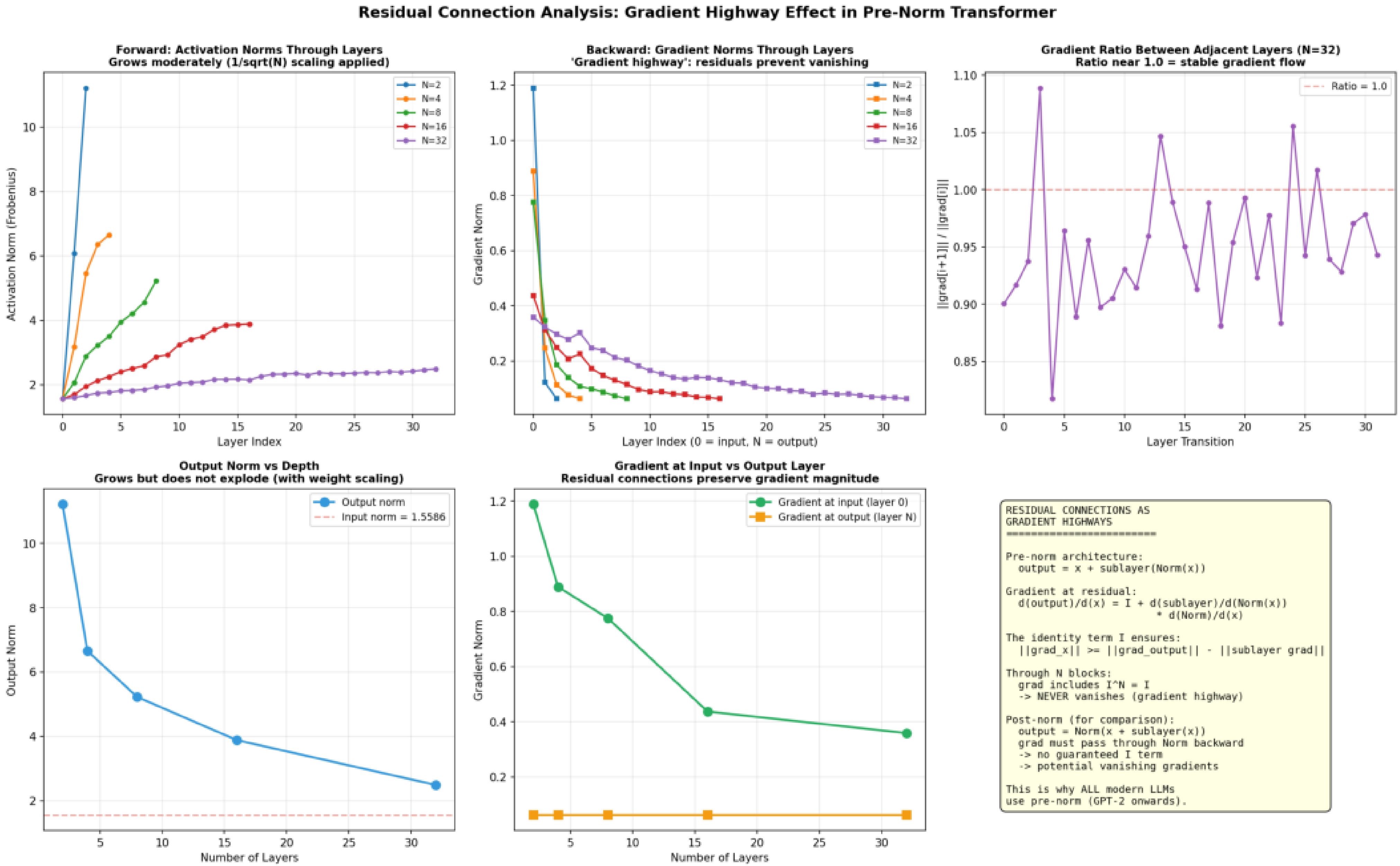
Example 2: Parameter Distribution Analysis



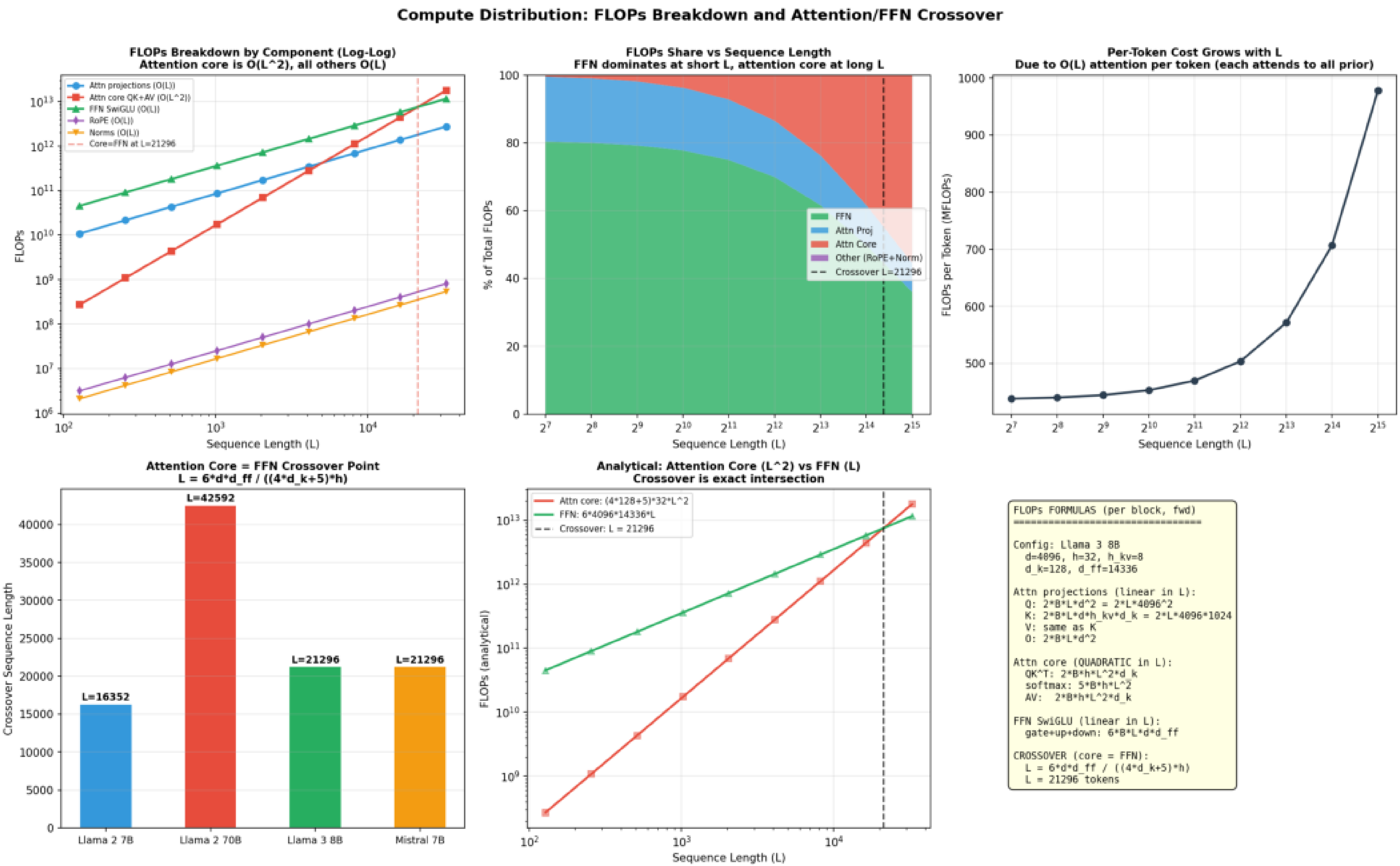
Example 3: SwiGLU Gating Visualization



Example 4: Residual Connection Gradient Highway



Example 5: FLOPs Breakdown and Crossover Analysis



Example 6: Causal Masking and Position Sensitivity