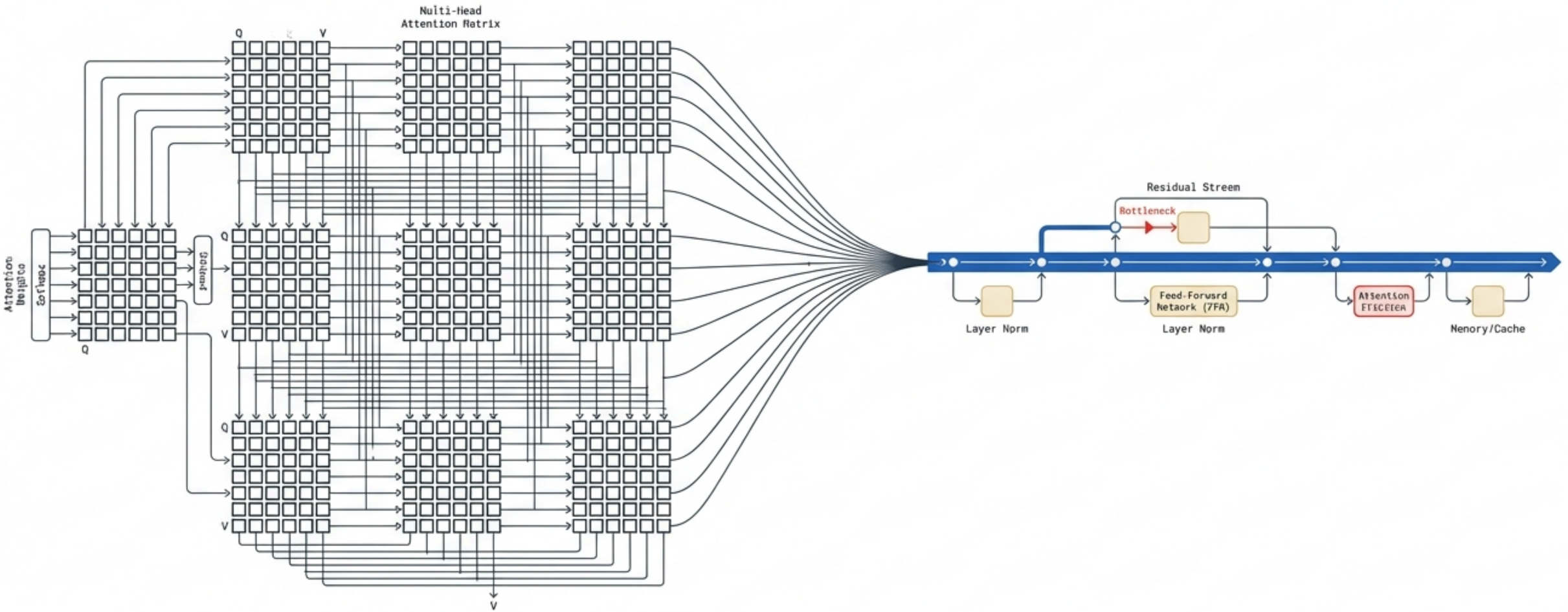
Deconstructing the Modern Transformer Architecture

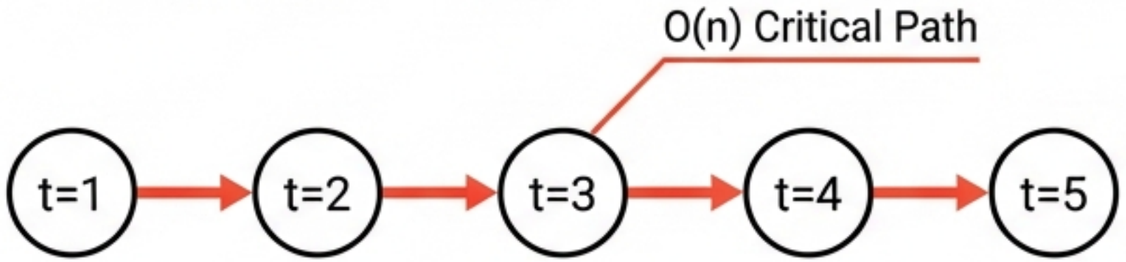
From canonical scaling laws to mechanistic interpretability and hardware-aware optimization.



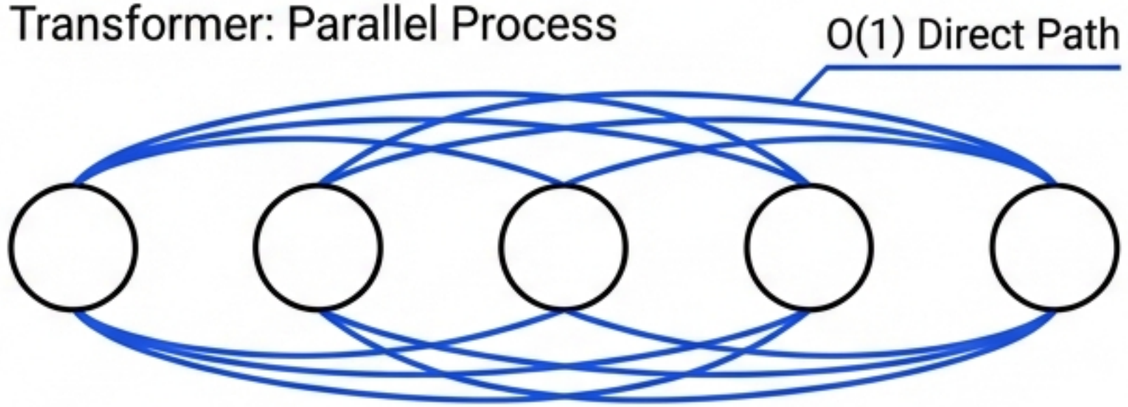
Eradicating the Sequential Bottleneck

The Transformer is a parameterized sequence-to-sequence architecture that eliminates sequential constraints, enabling total parallelization via an O(1) path length.

RNN: Sequential Process



Transformer: Parallel Process



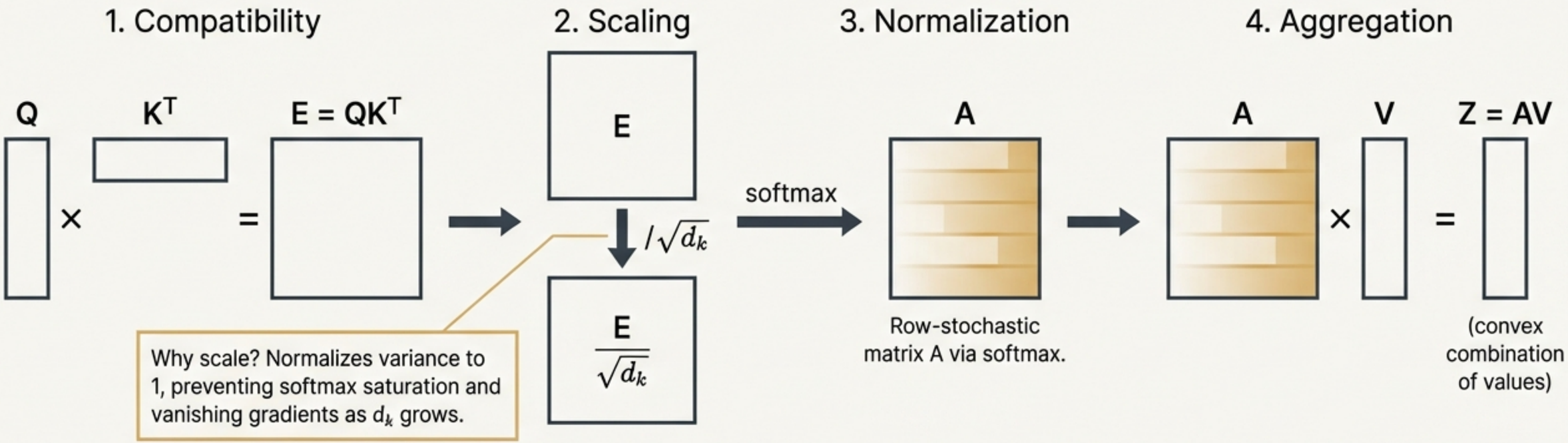
Property	RNN	Transformer
Sequential operations	$O(n)$	$O(1)$
Maximum path length	$O(n)$	$O(1)$
Computation per layer	$O(n \cdot d^2)$	$O(n^2 \cdot d + n \cdot d^2)$

The O(1) path length enables direct gradient flow and massive parallelism, introducing an $O(n^2 \cdot d)$ pairwise attention cost.

The Core Engine of Soft Dictionary Lookup

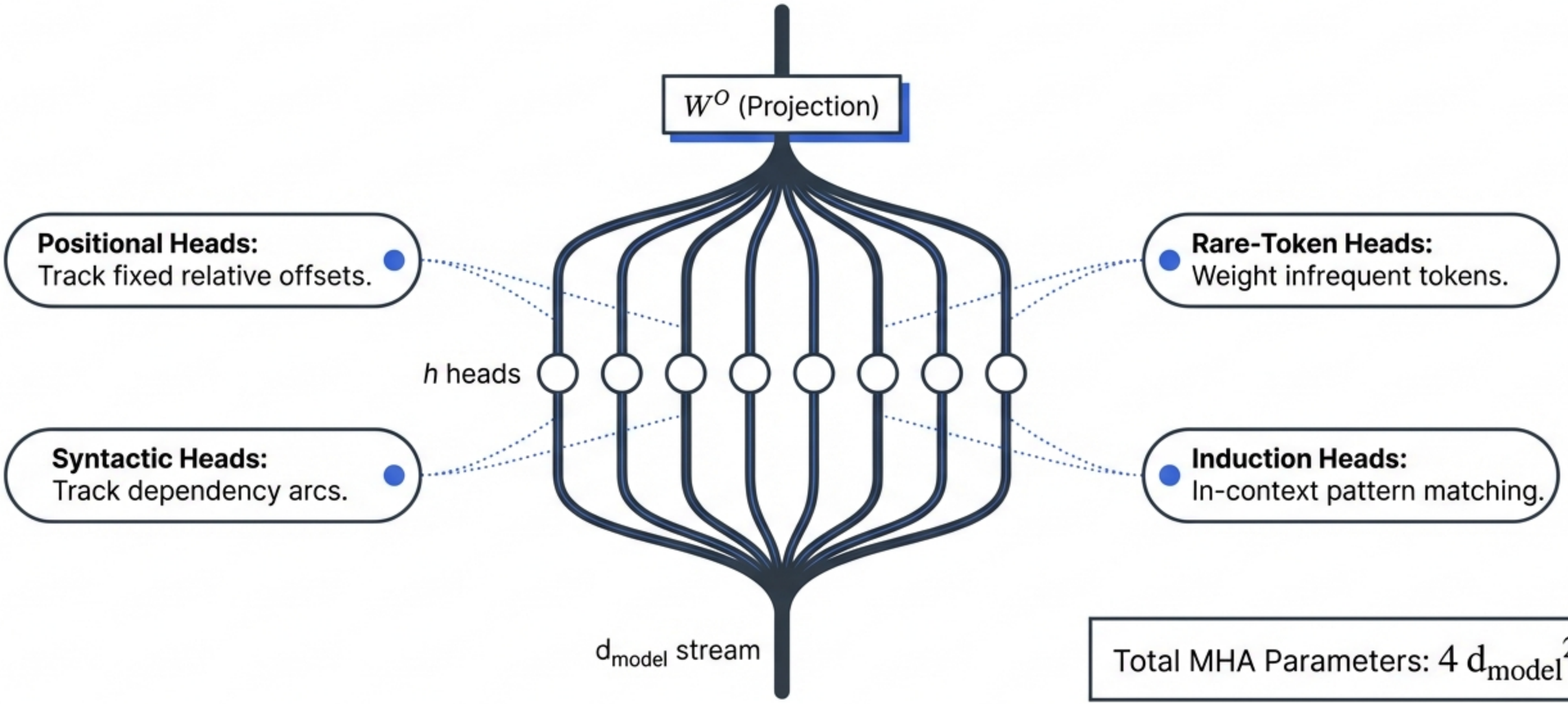
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\,\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V}$$

Q: What am I looking for? **K**: What do I contain? **V**: What do I provide if selected?



Attending to Multiple Representational Subspaces

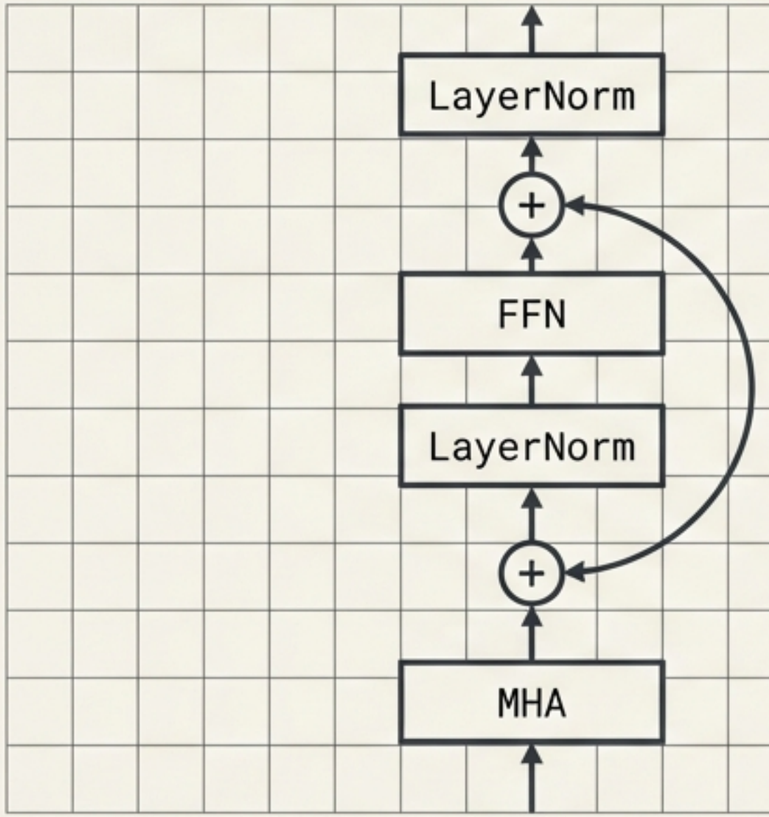
Multi-Head Attention (MHA) splits the residual stream to process distinct semantic features simultaneously.



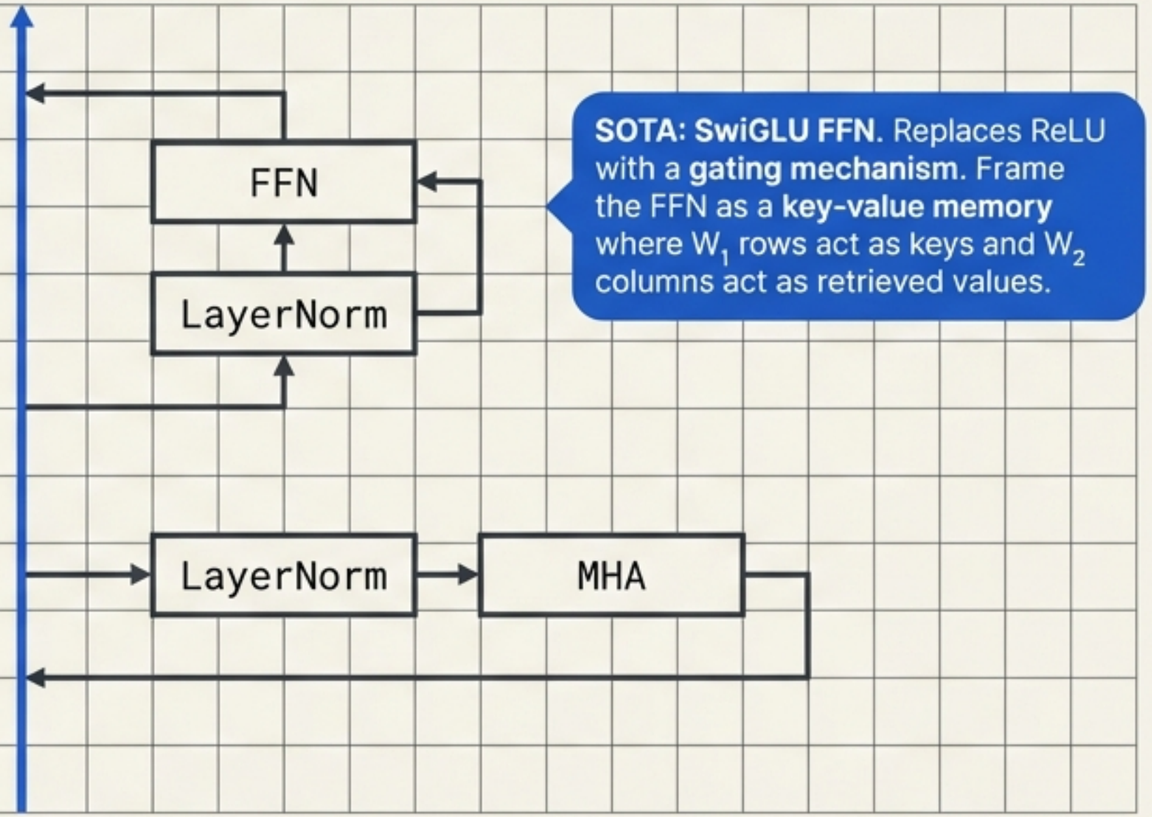
Total MHA Parameters: $4\,d_{model}^2$

Assembling the Modern Encoder Block

Original Post-Norm (2017)



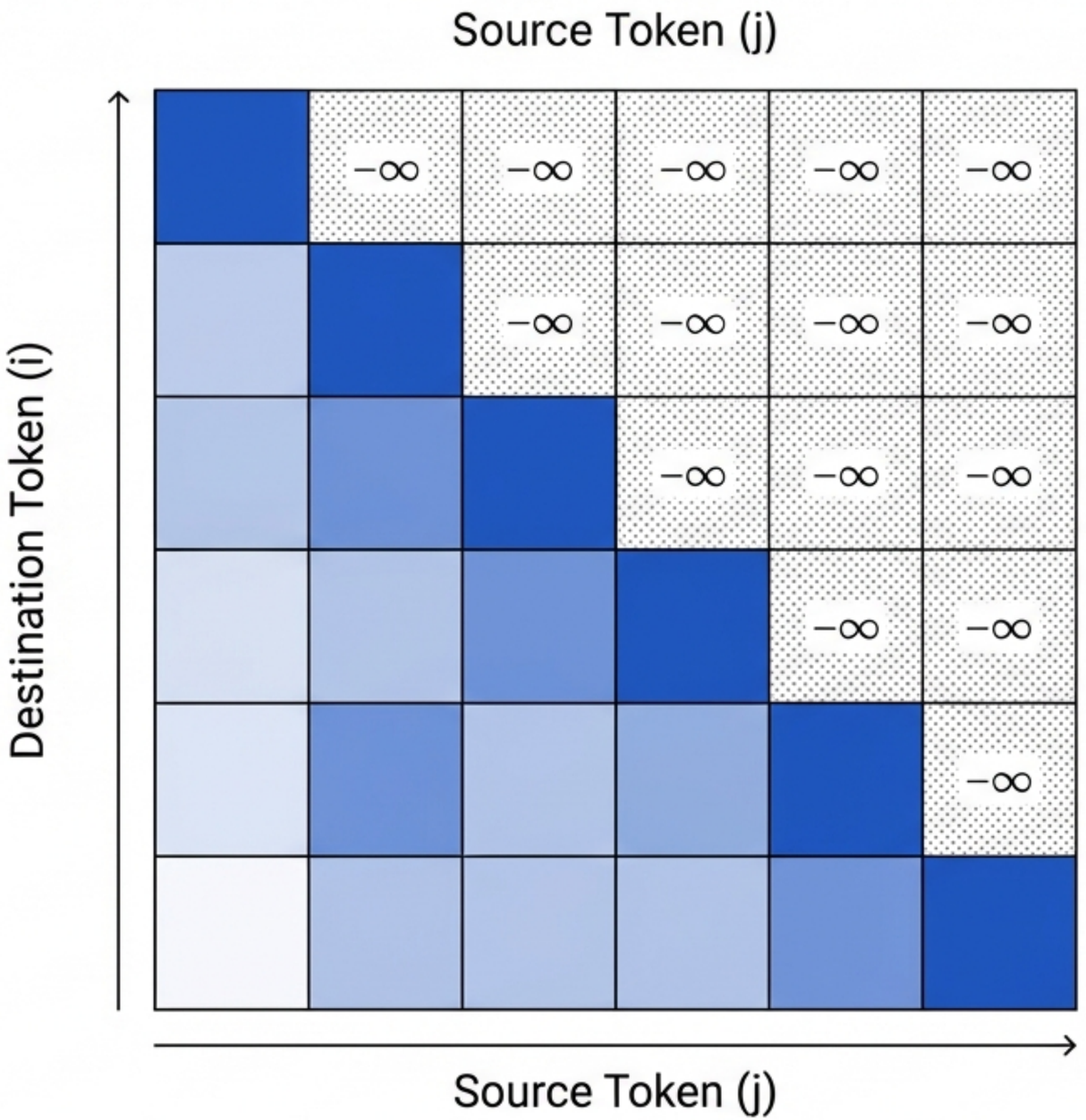
Modern Pre-Norm (SOTA)



SOTA: SwiGLU FFN. Replaces ReLU with a **gating mechanism.** Frame the FFN as a **key-value memory** where W_1 rows act as keys and W_2 columns act as retrieved values.

Pre-Norm enables stable training without warmup by keeping the additive residual pathway completely unobstructed.

LayerNorm normalizes across feature dimensions within a single sample (unlike BatchNorm), adapting to variable-length sequences.



Causal Masking Enables Autoregressive Generation

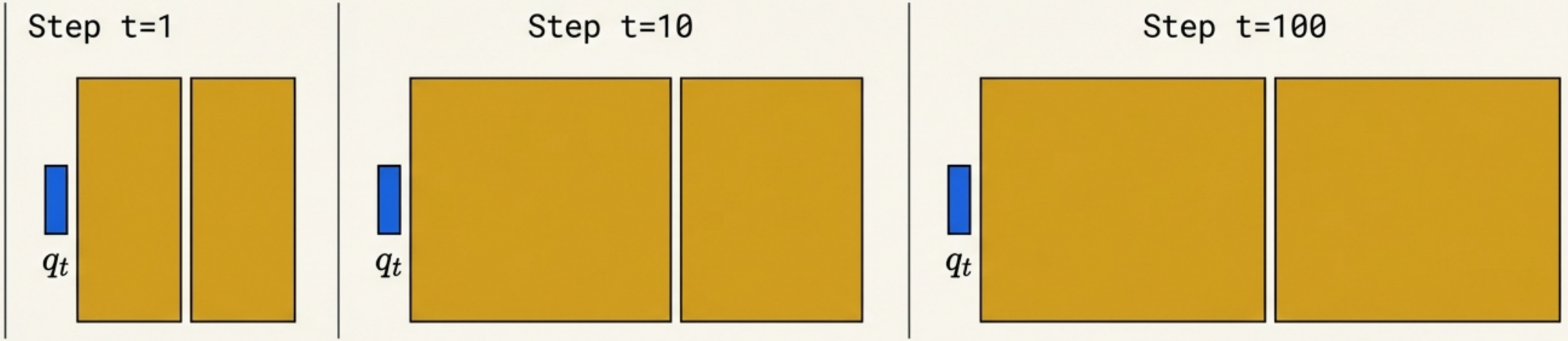
- Causal **masking** enforces the autoregressive property $P(y_t \mid y_{<t})$ by preventing position i from attending to $j > i$.
- This mechanism enables **parallel teacher forcing** during training despite the sequential nature of generation.

SOTA Badge

SOTA Shift: Decoder-Only Architectures (GPT, LLaMA). The prompt and generation are treated as a single causal sequence, unifying pre-training and generation.

The KV-Cache Exposes an Inference Memory Bottleneck

Autoregressive inference compute drops from $O(t^2 \cdot d_k)$ to $O(t \cdot d_k)$ by caching prior keys and values.



Dominant Memory Bottleneck Equation:

$$2 \cdot B \cdot L \cdot h \cdot n \cdot d_k \cdot \text{bytes\_per\_element}$$

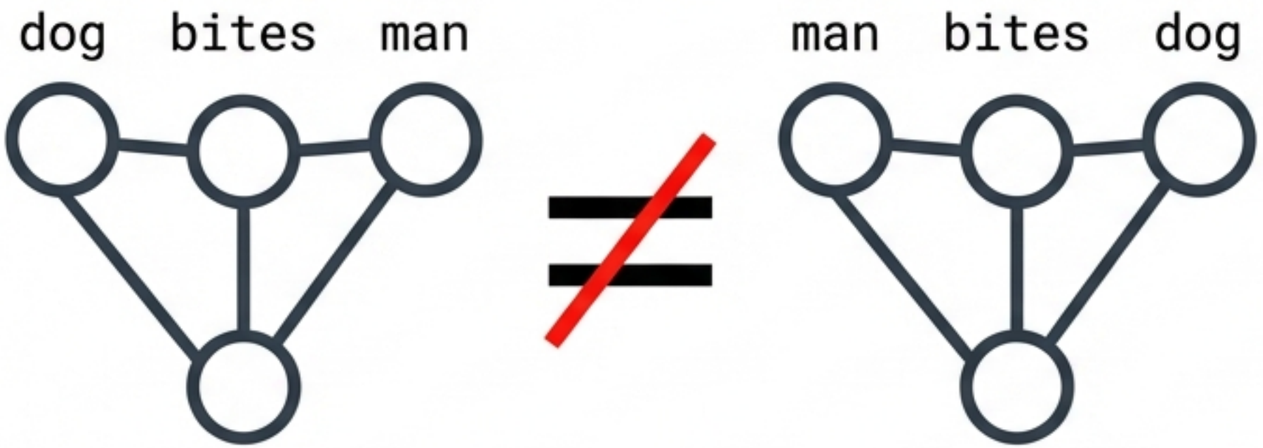
Compute is cheap, memory bandwidth is expensive. The KV-cache dictates maximum batch size and sequence length during deployment.

Breaking Permutation Equivariance with Spatial Context

The baseline attention mechanism has no inherent concept of sequence order.

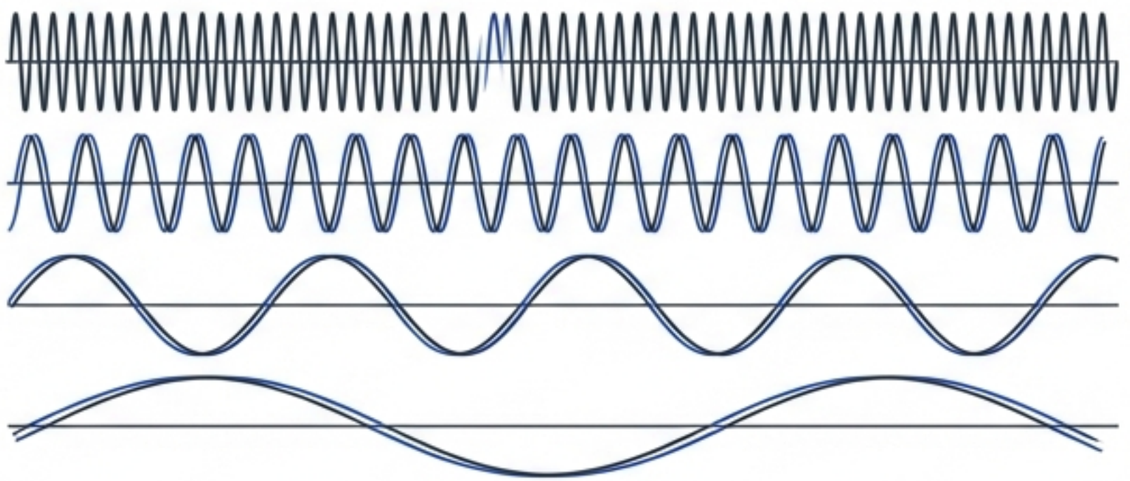
$$\text{Attention}(\prod \mathbf{X}) = \prod \text{Attention}(\mathbf{X})$$

Problem



Without spatial context, the model perceives these two sequences as identical despite the semantic difference.

Canonical Sinusoidal (2017)



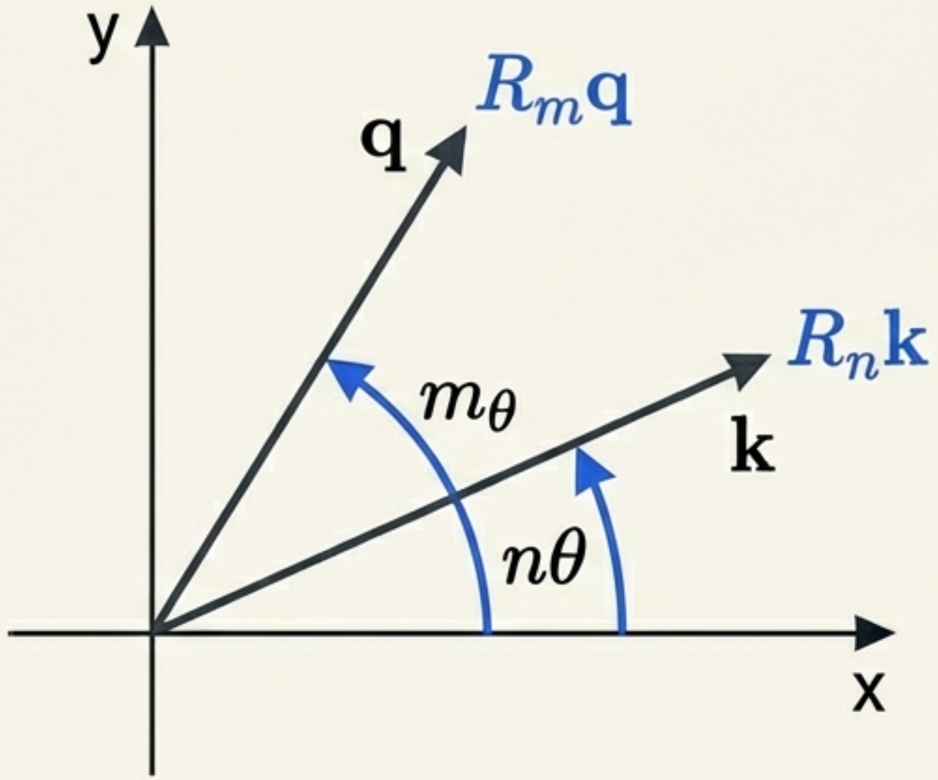
Uses pairs of sine/cosine functions to provide a geometric spectrum of wavelengths. Relative positions are derived via a linear transformation matrix.

Learned Embeddings

Uses fixed lookup tables W_{pos}. Fails to extrapolate beyond the maximum token length seen during training.

Modern Positioning Relies on RoPE and ALiBi

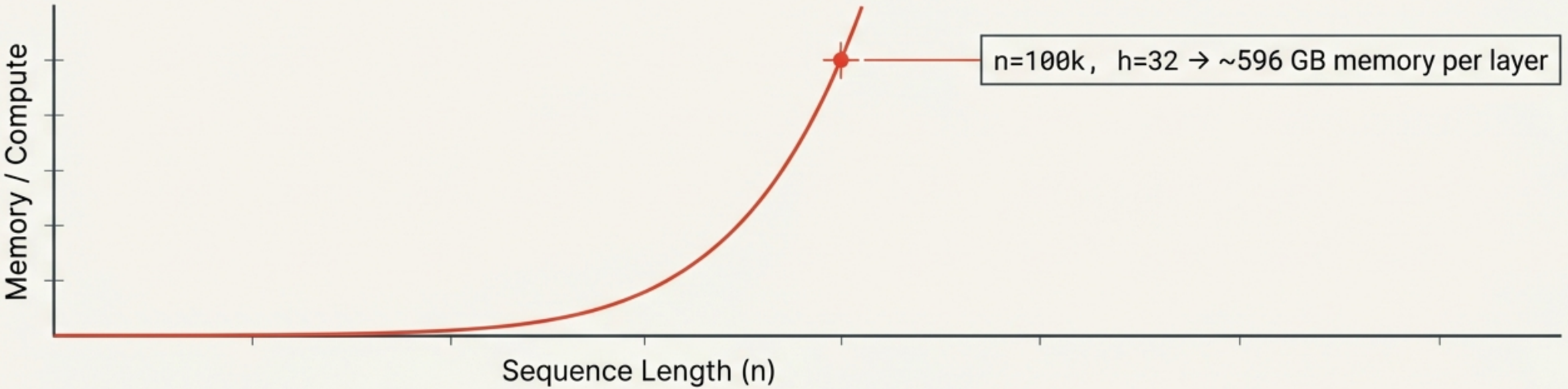
RoPE (Rotary Position Embedding)



Transforms attention scores via orthogonal 2D rotations. Injects explicit relative position directly into the dot product. Zero additional parameters.

ALiBi (Attention with Linear Biases)

Adds a static, head-specific linear bias directly to attention logits. Exponential distance decay, zero parameters, massive length extrapolation (100K+ tokens).



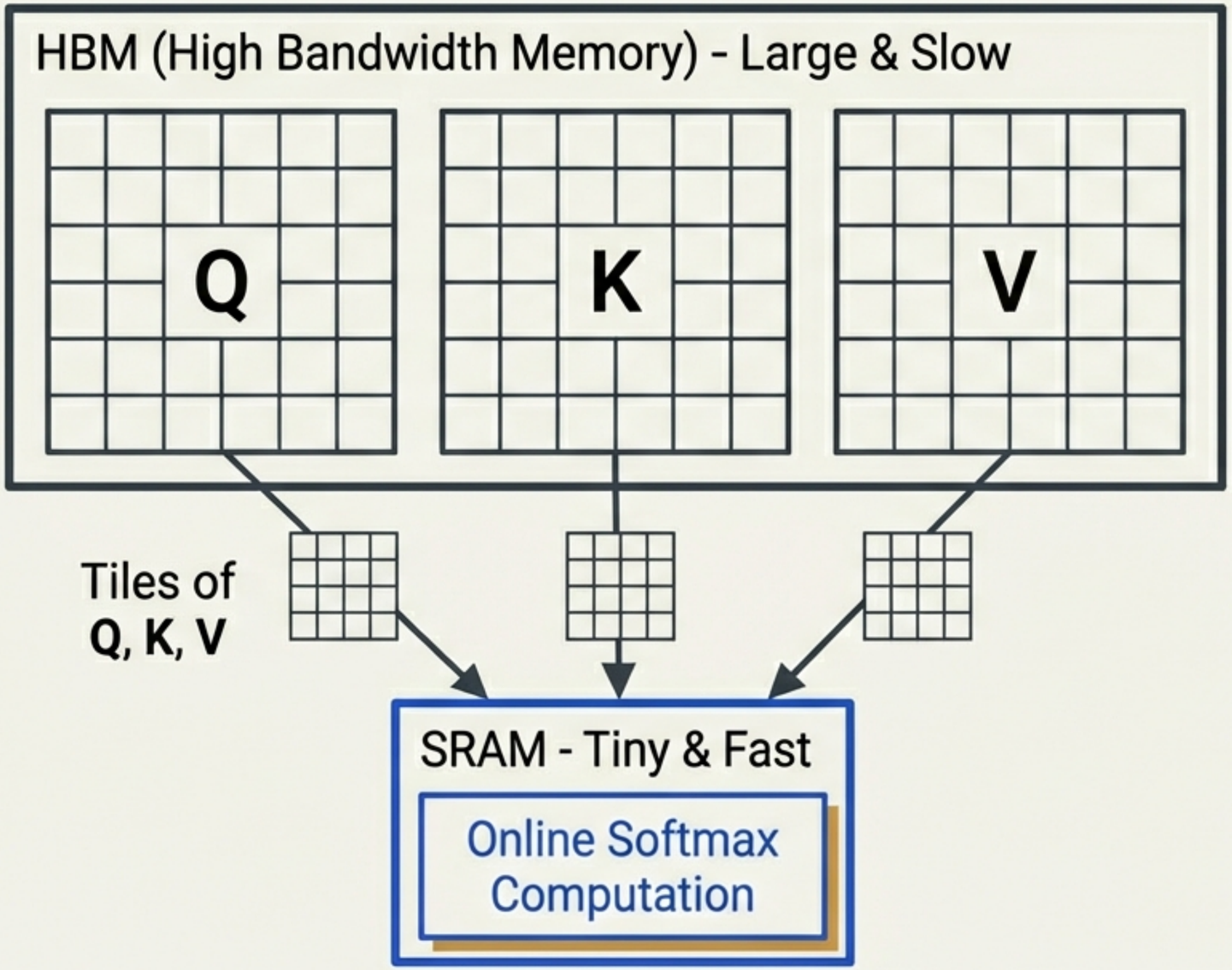
The Quadratic Scaling Wall at 100K+ Tokens

Materializing the $n \times n$ attention matrix demands $O(n^2)$ memory and $O(n^2 \cdot d)$ compute.

Early Algorithmic Solutions (Taxonomy):

- **Sparse Attention**: Sliding windows (e.g., Longformer) yield $O(n \cdot w \cdot d)$ complexity but lose global context without explicit global tokens.
- **Linear Attention**: Approximates softmax via kernel feature maps for $O(n \cdot D \cdot d)$ complexity, but precision degrades on sharp attention distributions.

Hardware-Aware Exact Attention via FlashAttention

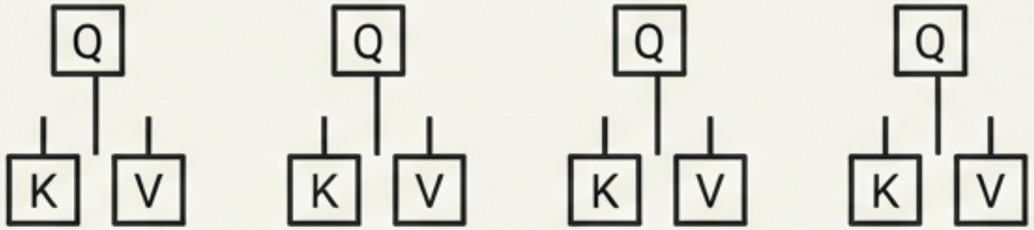


FlashAttention computes exact attention without materializing the $n \times n$ matrix in memory, utilizing an Online Softmax trick.

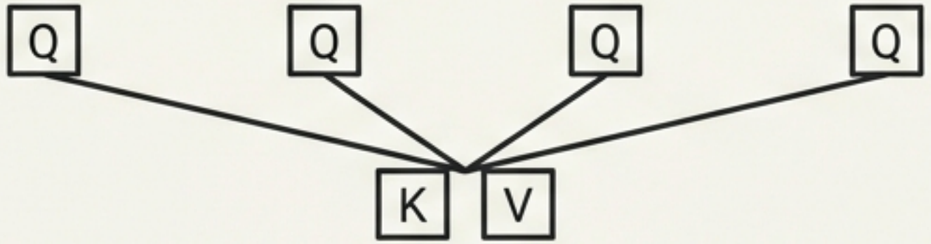
FLOPs: $O(n^2\, d)$ (Exact, no approximation loss)
Memory footprint: Reduced to $O(n)$
HBM Read/Writes: Slashed from $O(n^2)$ to $O(n^2\, d^2 / M)$

Compressing the KV-Cache with GQA and MLA

Multi-Head Attention (Baseline)

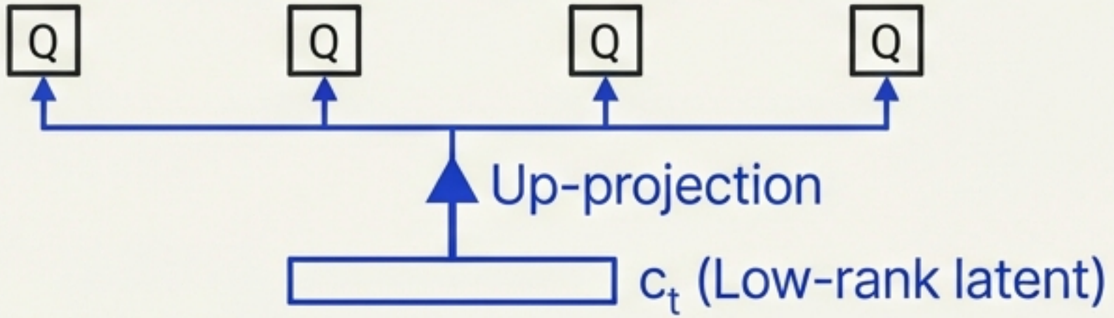


Grouped-Query Attention (LLaMA-2)



GQA strikes an optimal tradeoff: ~1.5x faster inference than MHA with minimal quality loss by grouping queries to share a single K, V pair.

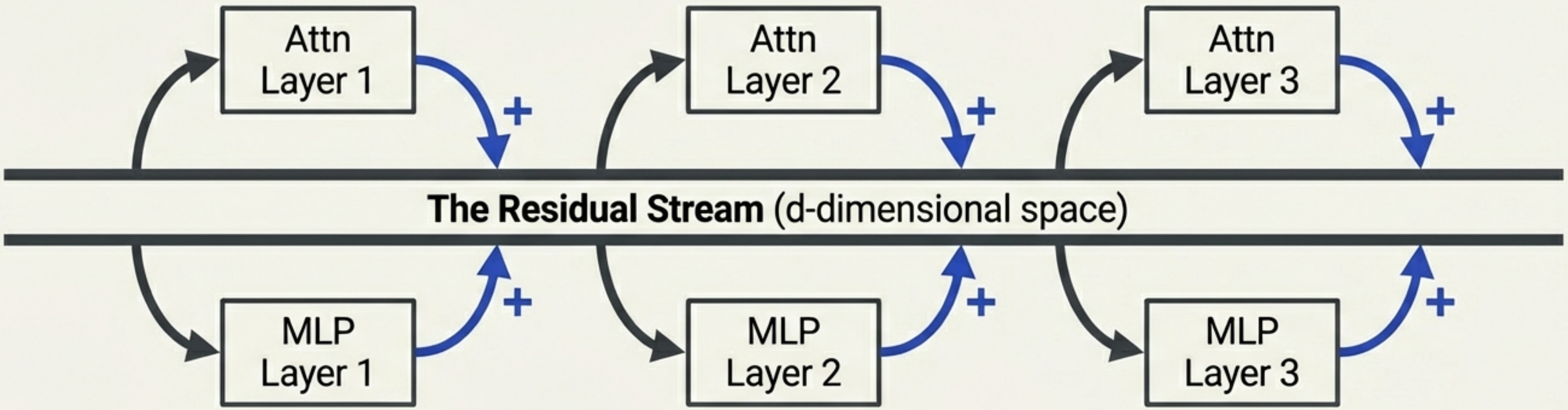
Multi-Head Latent Attention (DeepSeek-V2)



SOTA: MLA compresses K and V into a shared, extremely low-rank latent vector c_t, drastically shrinking the cache footprint while maintaining expressiveness via up-projection.

Reconceptualizing the Architecture as a Residual Stream

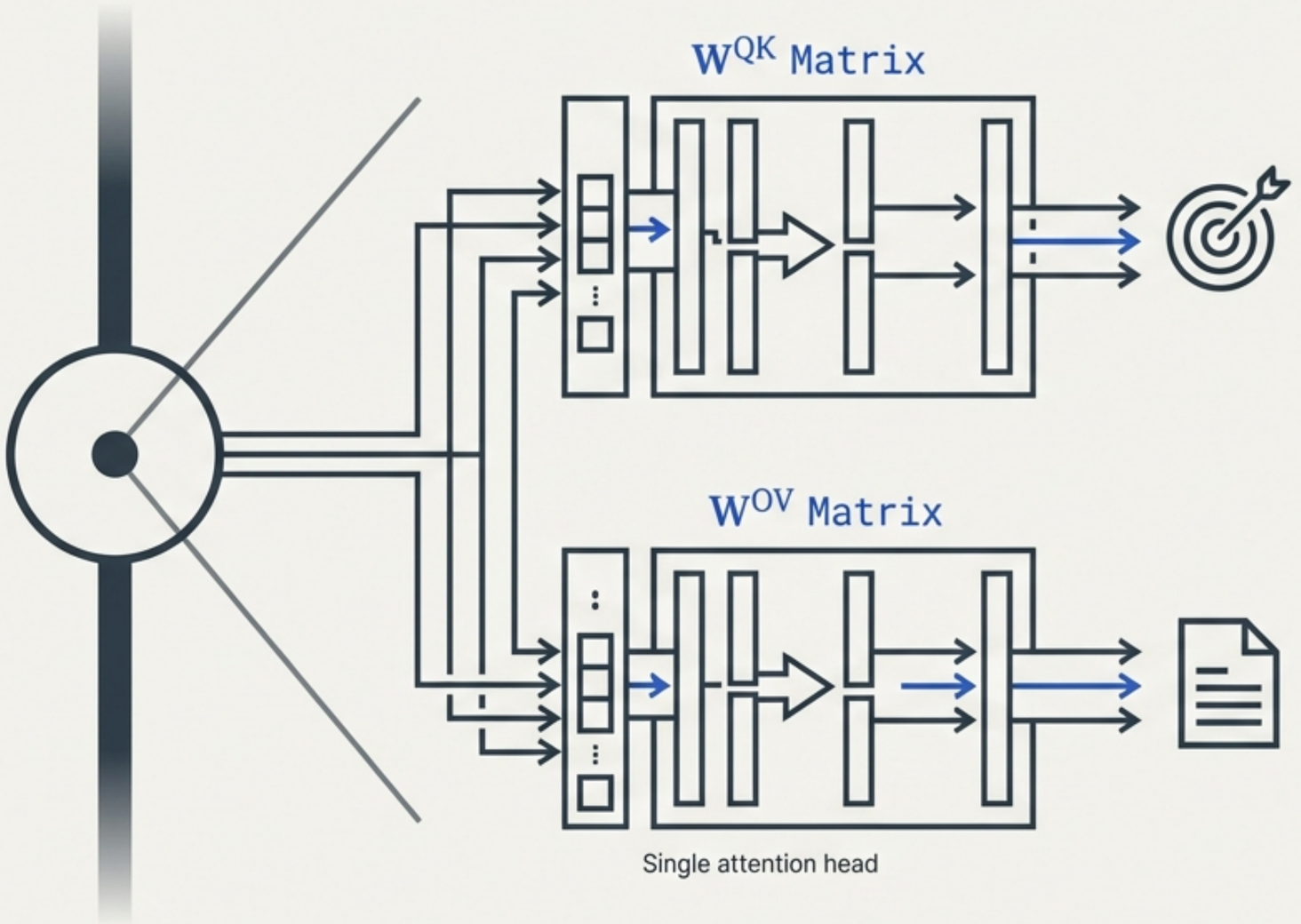
$$\mathbf{x}_{\text{final}} = \mathbf{x}^{(0)} + \sum \mathbf{a}^{(l)} + \sum \mathbf{m}^{(l)}$$



Communication Channel
Layers do not transform the hidden state. They read from a shared d-dimensional space and write corrective updates back to it.

Composition
Because it's a shared stream, later layers can read the specific output writes of earlier layers, creating compositional reasoning circuits.

Mechanistic Interpretability Opens the Black Box



QK Circuit ($W^Q W^{KT}$): A bilinear form determining **WHERE** to attend.

OV Circuit ($W^V W^O$): Determines **WHAT** information to write to the destination token.

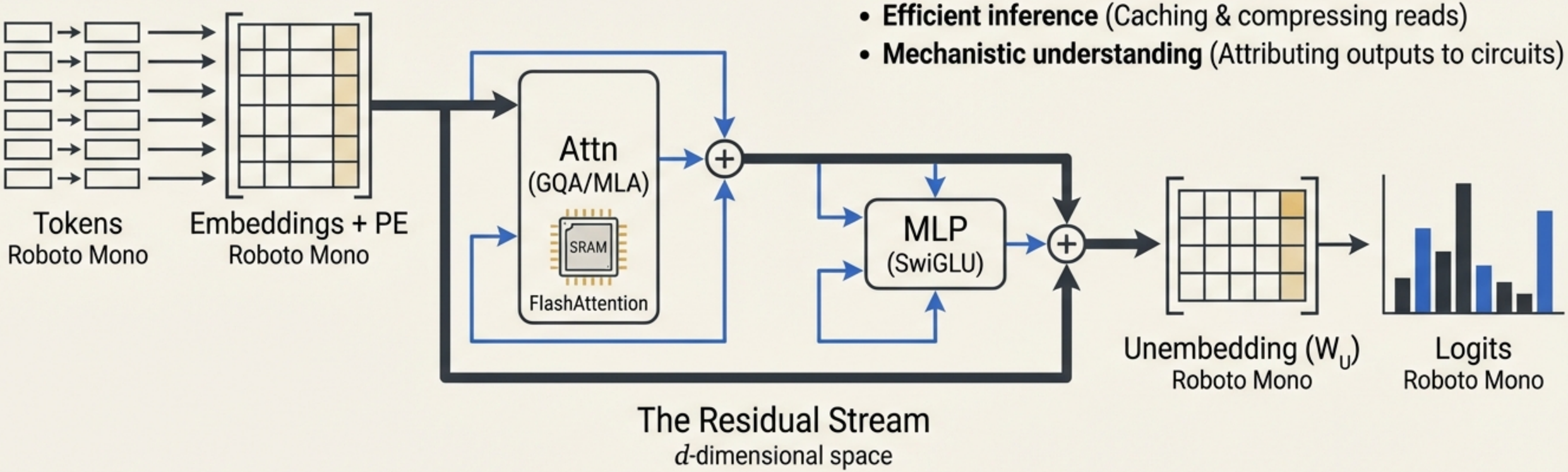
Direct Logit Attribution (DLA): Quantifies exactly how much a specific head promotes or suppresses a final output token logit by projecting its write directly via the unembedding matrix W_U.

Superposition & SAEs: Sparse Autoencoders extract compressed, nearly-orthogonal activations into distinct, monosemantic features.

The Complete Read-Process-Write Information Flow

The **Unifying View** elegantly unites:

- **Architecture design** (What writes to the stream)
- **Efficient inference** (Caching & compressing reads)
- **Mechanistic understanding** (Attributing outputs to circuits)



The Transformer is *not* a sequence of opaque transformations; it is a mathematically transparent, highly optimizable, fixed-width communication channel.